

Precaution versus principles

Constitutional law can make the eyes glaze over, but scientists need to be wary of subtle amendments that may undermine a fundamental liberty enshrined in most constitutions: scientific freedom.

A vote by France's National Assembly last week to write the 'precautionary principle' into the country's constitution is misguided. It is part of an environmental charter, which Jacques Chirac, the French president and new-found champion of the environment, hopes to give the same constitutional standing as the 1789 Declaration of the Rights of Man and the Citizen.

The precautionary principle has featured in environmental law since the 1970s, and appears in several international treaties, national laws and the proposed European Constitution. Invariably vaguely worded, it states that if technology or actions might cause harm, and there is scientific uncertainty as to what and how much, then other actions should be taken to anticipate and prevent that harm.

The devil in this sensible proposition lies in its interpretation. Wisely, most texts leave this to the checks and balances of legislators, whose deliberations are usually balanced and flexible. The French charter, in contrast, makes the precautionary principle "directly applicable", meaning that anyone can contest the desirability of a proposal even if it falls outside the scope of existing legislation.

The French Academy of Sciences has rightly warned that this may have "disastrous consequences". Pressure groups are sure to seize on it to block research they dislike. Parliamentarians should listen to these concerns as the bill moves through the approval process.

When European leaders meet in Brussels next week to try to agree

on the European Constitution, they will find a proposed amendment stipulating that the European Union (EU), in all its policies, "shall pay full regard to the welfare requirements of animals, as sentient beings". It is probably the least of the negotiators' worries, and may slip through without due consideration of the licence it would give to obstruct research that most people would support. The same wording on animal welfare was previously added to the EU's founding treaties in 1997. Given that a coalition of European animal-rights and welfare organizations have lobbied so strongly for its inclusion in the constitution, scientists should anticipate that giving the EU direct powers on animal welfare will, at the very least, result in yet stricter controls and greater bureaucracy.

Issues such as the environment, animal welfare and bioethics should not, in most cases, be addressed in constitutions. A battery of legislation on such issues already exists, ranging from laws passed by national parliaments to codes of conduct produced by professional bodies and recommendations by ethical committees, which although devoid of the power of law, often have equivalent impact or directly influence legislation. Unfamiliar, complex and fast-moving areas involving science can often be dealt with at this lower legislative level. Indeed, it is better placed to juggle regulations to meet each nation's own mix of religious, moral and cultural traditions, as well as the sometimes conflicting streams of contemporary thought. ■

Boosting bioprospects

A new approach to registering ownership should rebuild confidence, benefiting all stakeholders.

Exploring plants to find new compounds for drugs is an age-old practice. But as scientific sophistication in probing genetic resources has increased, the politics of the endeavour now called 'bioprospecting' have deepened to a point where the rate of discovery has been seriously eroded.

A path out of the forest of international diplomatic hazards must be found. A new phase of negotiations through the Convention on Biological Diversity may provide that course — and not a moment too soon, as economic pressures in developing nations are wiping out the very biodiversity that the convention was meant to save.

The convention — set up 12 years ago and now adopted by 188 nations — was designed as a conservation tool with an incentive system. It was intended to save wildlands and give developing nations, particularly in the tropics, a share of the rewards of bioprospecting.

But making this model work in the real world of suspicion and competing economic needs has been harder than many imagined. Research projects have been delayed or halted as nations have developed restrictive regulatory systems, spurred at times by political demagoguery. Drug companies have stopped funding bioprospecting enterprises, surely a short-sighted business strategy. As a result, few economic rewards are returning to developing nations.

Despite these shortcomings, now is not the time to abandon the convention's ideals. There have been some notable successes (see page 598). Researchers who operate transparently while building the

capacity to conduct research within developing nations are demonstrating that the convention's goals are achievable. And the importance of conservation worldwide has become increasingly plain.

After its seventh conference of the parties in February in Kuala Lumpur, Malaysia, the convention agreed that during the next two years an added framework for access and benefit sharing would be created. A draft of this framework is to be presented in Brazil in May 2006 at the next conference of the parties.

A key component of this framework is a certificate of origin that links compounds discovered in the jungle to drugs that may be developed at pharmaceutical companies. Such a document could eliminate developing nations' fears of a rip-off. And it could provide drug-company investors with a clear chain of a compound's ownership.

Convention delegates will meet in Bangkok early next year and in Spain in 2006 to construct a draft of this framework for the Brazil meeting. They should work to make the creation of certificates of origin a reality.

Non-governmental organizations obsessed with blocking the use of genetic resources should take a step back and look at the opportunities that discovery provides. Those economic powers that fear a challenge to the US patent system should see the new framework as an enhancement of property rights, not something to undermine them. And leaders of nations with the richest biological diversity should ensure that the new framework enhances conservation. ■





All change
Africa pressed to
switch malaria
drugs, despite cost
p588



Drug bust
Pharmaceutical firm
targeted by New York
attorney-general
p589



Lack of porpoise
Mexican fishermen's
nets endanger rare
marine mammal
p590



Fair funds
Economists urge
redistribution of
development aid
p592

Gene therapists hopeful as trials resume with childhood disease

Erika Check, Minneapolis

A French gene-therapy trial that cured nine children of a severe disease, but gave two of them cancer, looks set to restart after a 22-month suspension.

The trial involves children who suffer from severe combined immunodeficiency disease (SCID). These children lack innate defences against infections and without treatment they can only survive in isolated environments. One US gene-therapy trial for the disease has also restarted, and others are likely to resume this year.

The suspension of the trials had deeply shaken the gene-therapy field, because SCID was the only disease that had ever been cured by such therapy. Researchers at the annual meeting of the American Society of Gene Therapy in Minneapolis, Minnesota, last week saw the resumption of the SCID trials as a bright spot after a long dark spell for the field.

Specialists say there is still a risk that some children will develop cancer during the trials. But they say the trials should proceed, because the French technique has cured many children who suffer from the devastating illness.

"We're moving forward," says Donald Kohn, past president of the American Society of Gene Therapy and leader of one of the US trials. "No therapy is without risk, and now that we've had time to look back, we realize that this therapy even with the risk may be better than the current treatment," Kohn says.

The children in the French trial suffer from a version of the disease called X-linked SCID. For X-linked SCID patients, the alternative to gene therapy is a bone-marrow transplant. But these transplants are successful in only 70% of children, unless they have a suitable bone-marrow donor. Out of 18 children treated using gene therapy, 15 appear to have been cured of X-linked SCID.

The French trial, led by Alain Fischer of the Necker Hospital in Paris, was the first to show that infants could be cured through doses of a gene to correct their genetic deficiency. But in September 2002, Fischer announced that he had halted his trial because one of the participants had developed leukaemia. Another



Beyond the veil: Donald Kohn is one of those hoping to cure infants who are unable to fight infections.

child came down with leukaemia a few months later. Both are alive and recovering from their cancers.

Fischer's announcement prompted the US Food and Drug Administration (FDA) to stop three gene-therapy trials in America — two in X-linked SCID, and one in another form of the disease called ADA-SCID. But an X-linked SCID study in Britain was allowed to continue, and seven children have now been treated in that study. Claudio Bordignon of the San Raffaele Telethon Institute for Gene Therapy in Milan, Italy, was also allowed to treat patients with ADA-SCID for whom other therapies had failed. Five patients have been treated in that trial, and not one has developed cancer.

The first SCID study to be resumed in the United States is led by Harry Malech and Jennifer Puck of the National Institutes of Health at Bethesda, Maryland. They were cleared to begin their trial in December and treated one child with X-linked SCID in January. So far, his condition is stable, Malech says.

The other two US trials are led by Kohn and by Kenneth Weinberg, both of the Children's Hospital Los Angeles. Both say that they are consulting with the FDA and hope to resume their trials later this year.

Since 2002, scientists have learned more about why gene therapy caused the X-linked SCID patients to get cancer. Such patients receive a copy of a gene they lack, called the *gamma-C* gene. This gene allows their immune cells to grow normally. But in the children who get leukaemia, *gamma-C* seems to switch on a cancer-causing gene called *LMO2*, which is found in human DNA.

Fischer and other scientists will adjust their treatment plans to minimize risks from this switching effect. For instance, in most cases Fischer will now only treat children older than 6 months, because they might be less vulnerable to cancer than the very young babies who developed cancer in his trial. Fischer will also place an upper limit on the number of corrected cells he injects into the children.

Some researchers take a different tack when balancing the risk of cancer against potential cure. Weinberg is asking the FDA to allow him to resume his X-linked SCID trial without limiting the age of the children enrolled or the dose of cells they receive. But like all other leaders of SCID trials in the United States, Weinberg will monitor each of his patients for signs of cancer for at least a decade after the trial.

Global fund changes tack on malaria therapy



African patients currently receive drugs to which malaria is becoming resistant.

Declan Butler, Paris

An outcry from researchers has helped prompt a move to switch malaria treatment in Africa to a more effective therapy. The drug, artemisinin, comes from a Chinese herb, and cures 90% of patients in three days.

But aid donors have shrunk from promoting artemisinin, largely because it costs at least ten times more than established malaria treatments, to which the parasite that causes the disease has grown resistant.

At a closed meeting last month, the Geneva-based Global Fund to Fight AIDS, Tuberculosis and Malaria agreed to instruct African countries to retrospectively modify all malaria grants awarded to specify only the newer drug. The move will cost the fund more than a billion dollars over the next five years, says Vinand Nantulya, one of its senior officials.

All future funding will stipulate the use of artemisinin. The policy change is expected to force most countries to change their national drug policies. "Money talks," says Allan Schapira, coordinator of malaria policy and strategy at the Roll Back Malaria team of the World Health Organization (WHO).

Nantulya admits that the move was given "further impetus" by the discussion that followed publication of an article in *The Lancet* in January in which prominent scientists accused the fund and the WHO of "malpractice" for bowing to pressure from aid-giving countries. They also cautioned against the use of outdated drugs (*Lancet* 363, 237–240; 2004).

Donors such as the United States and Britain had discouraged African countries from adopting artemisinin on the grounds that it cost too much, that data on its effect in children were insufficient, and that it was not

needed in regions where other anti-malaria drugs still worked, says Amir Attaran of the Royal Institute of International Affairs in London, and the article's lead author.

In 2002, the WHO recommended artemisinin as the treatment of choice, but in Africa only Zambia and Zanzibar have adopted it.

The growing resistance of the malaria parasite *Plasmodium falciparum* to chloroquine and sulfadoxine-pyrimethamine, the mainstay of treatments in Africa — where 90% of global malaria deaths occur — has also damaged the lofty ambitions of the Roll Back Malaria initiative. This was launched in 1998, with the goal of halving malaria deaths by 2010.

But deaths are up, not down, and mortality rates of treated patients are doubling in many parts of Africa. John Lidén, a spokesman for the Global Fund, says that it revisited resistance data from the WHO and other sources, and that this "composite set" convinced it to act. "We are results-based," he says, "and if we find we are paying for medicines that don't work, we want to pay for those that do."

The higher cost of artemisinin means that African countries will spend their revised grants more rapidly than they had intended. The drug costs about US\$2 for each adult treatment, compared with ten cents for existing drugs. Its price is expected to fall to about \$1 with improvements in production.

Countries might be allowed to apply for further grants in advance to make up the loss, Lidén says, and donors may be asked to release an additional \$1 billion to pay for the drug. "I appeal to the donor community to recognize that moving countries to using the right drugs will require much more money," says Nantulya, "They must take this seriously." ■

Biotech industry struggles with generics approval

Jonathan Knight, San Francisco

Leaders of the biotechnology industry hit out this week at the slowness of regulators in getting to grips with a problem that is confronting the industry as it comes of age: how to get approval for generic versions of biotech drugs after patents expire.

At the annual meeting of the Biotechnology Industry Organization (BIO) in San Francisco this week, industry managers and lawyers expressed frustration at not knowing what they have to do to get generics approved. "We don't know what the rules are," complains Meredith Manning, a regulatory lawyer at Washington, DC-based law firm Hogan & Hartson, who represents biotech firms. This will considerably

lengthen the time for cheaper generics to reach the market, she says.

Drug patents are now expiring 17 years after they were first filed. In the United States, generics need approval from the Food and Drug Administration (FDA). This is straightforward for most drugs, which have a single active ingredient — usually a small molecule — whose purity and composition can be confirmed by chemical testing.

But the drugs that the biotechnology industry has developed tend to be large molecules such as antibodies and other proteins, whose identities cannot be confirmed by a simple chemical analysis. Manufacturers of generic versions of these products will need to show equivalence with

the original in another way. And the FDA has not yet said how this should be done.

Biotech generics are on sale now, but they have had to submit to costly new drug reviews, including clinical trials. Among the generics approved by this route are recombinant human insulin and the anti-cancer drug interferon-alpha. But dozens more biotech patents will expire in the next few years.

Janet Woodcock, director of the FDA's Center for Drug Evaluation and Research at Rockville, Maryland, says that the agency is working hard to have guidelines available for public review, perhaps later this year. "We fully intend to have a transparent process," she says. ■

Spitzer sues drug giant for deceiving doctors

Meredith Wadman, Washington

Eliot Spitzer, the politically ambitious attorney-general of New York state who has made his name pursuing fraudsters on Wall Street, has found a new target: drug companies who do not publicize negative research results.

On 2 June, Spitzer's office sued GlaxoSmithKline (GSK) for allegedly suppressing negative results of trials that tested the safety and efficacy of Paxil, its leading anti-depressant drug, in young people.

Spitzer's action is a novel and — for the pharmaceutical firms — potentially troubling incursion into a realm normally regarded as the province of the federal Food and Drug Administration (FDA).

He alleges that British-based GSK flouted New York's tough consumer-fraud law by "misrepresenting, concealing and otherwise failing to disclose to physicians" the negative results of four studies. These reported that Paxil, also known as paroxetine, was no more effective than a placebo for depressed young people, Spitzer's suit says, and that it might even increase the risk of suicide in this group.

Not only did the company fail to warn doctors of Paxil's risks, the suit contends, it promoted the drug to physicians for use in children and adolescents. The suit cites a 1998 internal company memo which said that in light of the poor study results, the company's "target" was to "effectively manage the dissemination of these data in order to minimize any potential negative commercial impact". It says that the company also gave its salespeople a memo stating that "Paxil demonstrates remarkable efficacy and safety in the treatment of adolescent depression". Attached to the memo, the suit says, was the one published study that showed some positive results for Paxil in the treatment of paediatric depression.

"GSK misled and deceived physicians and consequently the patients who relied on their professional judgment," the suit alleges. It estimates that Paxil was prescribed to some 2.1 million US youngsters in 2002, earning about \$55 million for the company that year. The drug lost US patent protection in 2003 and now faces stiff generic competition.

Tom Conway, the chief lawyer on the case in Spitzer's office, says: "It is fraudulent practice for a pharmaceutical manufacturer to release positive information while suppressing negative information. The facts of this case illustrate this principle perfectly, in our minds."

But GSK's chief executive, Jean-Pierre Garnier, immediately denied the charges. He told the *Wall Street Journal* that the US legal system is "out of control" and "is becoming an outrageous cost of doing business".

The company said in a statement that the 1998 memo "is inconsistent with the facts



On a mission: New York's attorney-general Eliot Spitzer (above) has set his sights on the pharmaceutical firm GlaxoSmithKline (left).

and does not reflect the company position". The statement says GSK has acted responsibly: "All paediatric studies have been made available to the FDA and regulatory agencies worldwide." Spokeswoman Mary Anne Rhyne added: "Data from paediatric studies of Paxil have been communicated publicly in more than two dozen presentations at major medical meetings."

Paxil has been approved by the FDA for use in adults with depression — but not children and adolescents. In Britain the drug is marketed as Seroxat, and in June 2003 the Department of Health, concerned about suicide risks, prohibited doctors from prescribing it to children under 18. The same month, the FDA recommended that it should not be prescribed to depressed young people.

However, if a drug has been approved for one condition, American doctors may then prescribe it for any complaint in any patient; this is called "off-label" use. Federal law does not require companies to make any studies of such off-label uses available to doctors. Spitzer's office instead charges GSK with consumer fraud, under state law.

Some doctors contend that the issue that really needs to be addressed is that neither the

companies nor the regulator have a well defined responsibility to publicize negative results. "It's perhaps too common for negative trials not to be submitted for publication," says Wayne Goodman, a professor of psychiatry at the University of Florida who sits on the FDA panel that evaluates psychiatric drugs. "That's not necessarily a problem if a drug is not going to make it to market. But if it is a marketed compound as in this case, then I think clinicians need access to all the available data."

David Fassler, a psychiatrist specializing in paediatric depression at the University of Vermont, adds: "If 20 studies are done, one is likely to show significance by chance. If that's the only one that gets published, and I don't know that there were 19 other studies with different results, it really skews the perception and ultimately alters clinical care."

The GSK lawsuit opens a new front in a wider war that Spitzer has declared against corporate targets. The 44-year-old Harvard Law School graduate, who became attorney-general in 1999, is widely believed to be preparing to campaign for governor of New York, and the cases he has launched have generated nationwide publicity. Spitzer has gone after the mutual fund industry, polluting power plants and the former New York Stock Exchange chairman Dick Grasso, whose "excessive" \$188-million compensation package Spitzer is challenging. And the current lawsuit may not be the last time the drug companies hear from Spitzer's office. "It's certainly an issue that is not going to end with this case," says Conway.

B. MATTHEWS/AP

GSK

Net losses pose extinction risk for porpoise



In peril: at least six vaquita have died this year after being accidentally caught in fishing nets.

Rex Dalton, San Diego

One of the world's most endangered marine mammals is dying in fishing nets at an alarming rate in the Gulf of California off Mexico, prompting fears that the species could soon be extinct.

Mexican scientists have confirmed that at least six of the porpoises, called vaquita (*Phocoena sinus*), have died in nets in the past six months. This level of reported deaths by fishermen is unprecedented, scientists say — and many more may actually be dying.

"At this rate, the species will not survive very long," says marine mammalogist Lorenzo Rojas Bracho, coordinator of Mexico's National Marine Mammal Programme in Ensenada.

The vaquita is now only found in the Gulf of California. Its population has dwindled to less than 600 animals after years of being accidentally snagged in the nets of large trawlers and small boats called pangas as they fish for shrimp and fish. The porpoises have been considered critically endangered by Mexican and international agencies for a decade.

In 1993 the Mexican government established a reserve area, including a no-fishing zone, in the upper Gulf, where fresh water trickling into the Gulf from the Colorado River is thought to provide ideal conditions for fish that the vaquita feeds on. But Mexican officials and scientists say that commercial shrimp trawlers and pangas

are continuing to fish the reserve illegally.

Fishing-industry officials and some members of the Mexican government blame the disappearance of the vaquita on US-owned dams in the Colorado River, which limits the amount of fresh water flowing into the estuary and may cut down the vaquita's food supply. But marine scientists who monitor the vaquita are unanimous that fishing nets are the main culprit. "Fishing clearly is the greatest threat," says Jay Barlow, a marine mammalogist at the Southwest Fisheries Science Center in La Jolla and a member of the recovery team.

Environmental groups including The Nature Conservancy, the WWF and Conservation International have devised a viable plan to save the porpoises, says Exequiel Ezcurra, president of Mexico's National Institute of Ecology in Mexico City. They propose that the Mexican government divert US\$50 million, currently used to subsidize fuel prices for the shrimp trawling industry, to instead purchase trawlers and decommission them. This would leave more fish for the panga fishermen, who would be encouraged to use fish traps and other methods that do not endanger the vaquita as gill nets do.

Ezcurra says the plan has not been implemented because the trawling industry won't give up its ships or make long-term commitments to halt fishing in the area. Industry groups did not respond to interview requests from *Nature*.

"I personally feel that the plan is the only ray of hope for the vaquita," says Ezcurra. "I am pushing for it very strongly." ■

Critics blast 'premature' paper on adult stem cells

Quirin Schiermeier and Martin Leeb, Munich

A stem-cell paper published last month in an online physics journal has created quite a stir. The paper, which was described as a "breakthrough" in adult stem-cell technology at a public announcement in Germany, is scientifically premature, experts say, and may lead to a surge in protests against research using embryonic lines.

In the current issue of *Applied Physics A* (doi:10.1007/s00339-004-2816-6; 2004) — an online journal published by Springer that normally reports materials-science research — a team of German biologists describes a technique for isolating stem cells from the pancreas of adult humans and rats. In culture, the cells differentiated into multiple cell lineages, such as brain and muscle cells, the team reports. If confirmed, the discovery would allow researchers often to use stem cells from adults instead of from embryos — a

particularly controversial source in Germany.

At a press conference in Berlin on 28 May, politicians of the Christian Democrats and Social Democrats referred to the study as "groundbreaking". Wolfgang Wodarg, health spokesman for the Social Democrats, said it was a blow to supporters of research involving human embryonic stem cells.

But leading stem-cell researchers rebuked the team and its host institutions — the University of Lübeck and the Fraunhofer Institute for Biomedical Engineering (IBMT) in St Ingbert — for publishing premature results. They point out that the paper was accepted one day after it was submitted, without full peer review. The paper provides little evidence for many of its claims, says Hans Schöler, a stem-cell researcher at the University of Pennsylvania in Philadelphia.

Michael Stuke, editor-in-chief of *Applied Physics A*, says he took the advice of several

experts before deciding to publish. "If I think that a topic is really hot and needs to be disseminated without the delay typical for major journals, I won't hesitate to invite biologists to publish in our journal," he says.

Publishing the data in an unreviewed physics journal was a "less-than-ideal solution", admits Günter Fuhr, director of the IBMT, who supervised the study. But he backs their decision to inform the public of their results at an early stage. Requests from *Nature* to interview the lead author of the paper, Charli Kruse of the University of Lübeck, were deferred to Fuhr.

Others say the paper's high profile will strengthen opposition to embryonic stem-cell research. "I am concerned that the whole field might suffer if news about alleged breakthroughs is spread in this way," says Oliver Brüstle, head of reconstructive neurobiology at the University of Bonn. ■



Failing to deliver: the ambitious Center of Advanced European Studies and Research in Bonn.

Futuristic centre under fire from German science council

Quirin Schiermeier, Munich

An ambitious interdisciplinary research centre, established in Bonn to do innovative work in a futuristic setting, has been found wanting by Germany's science council, the Wissenschaftsrat.

The Center of Advanced European Studies and Research (Caesar) opened in the former West German capital in 1999, with a generous, government-funded endowment of €380 million (US\$470 million). It was intended to strengthen science in the region and compensate its host city for the government's move to Berlin.

Caesar, whose 110 scientists last year moved into a chic, €60-million laboratory, focuses on the development of technologies and products relevant to industry, based on nanotechnology, biomedical engineering and advanced software design.

But the centre has so far failed to meet expectations, says a review by the council published on 1 June. It finds that too few marketable applications have been developed, and there are not enough publications, patents and spin-offs. Almost half of the centre's 21 groups are underperforming, and the quality-control system is less than satisfactory, the council's external reviewers found.

"Caesar's interdisciplinary concept is certainly worthwhile," says Wedig von Heyden, the council's secretary-general. "But in practice, research gets bogged down. To get back on track, the centre needs to focus on a reduced number of activities — preferably in the life sciences and medical technology."

Successful technologies developed at Caesar so far include a prototype biosensor for environmental analysis, and imaging software that lets oral surgeons design jaw implants in real time during an operation.

But many of the centre's group leaders lack the combination of business and scientific skills needed to find marketable products, says Gerhard Wegner, a scientific director at the Max Planck Institute for Polymer Research in Mainz, who chairs Caesar's scientific advisory committee.

"Caesar's concept and ambitions were over-optimistic from the start," Wegner says. "To succeed in these difficult markets, you need detailed knowledge about your industrial and scientific competitors, and a solid international standing. But a careful market analysis has never been done, and Caesar is more or less unknown outside Germany." He adds that he is not surprised by the outcome of the review. "The scientific advisory committee has mentioned these problems several times, but we didn't really get heard."

Karl-Heinz Hoffmann, Caesar's scientific director, denies that such advice has been ignored. "Cooperation between the different committees could perhaps have been better," he concedes. But he says he is "disappointed and annoyed" by the amount of criticism now descending on the centre.

Hoffmann says that some of his groups have only been working for three months, and that it is far too early to judge their performance. "We may have been a bit optimistic about the speed of development in, say, nanotechnology. But that is no reason to brush us aside so harshly."

The Wissenschaftsrat has recommended that Caesar's foundation board appoints an independent commission to restructure the centre by the end of next year. In the meantime, it says, no more money should be invested in the centre without the express approval of its grant-givers.

■ www.caesar.de

Nanotech takes small step towards burying 'grey goo'

Jim Giles

Nanotechnology researchers are sick of hearing about 'grey goo'. Their research is still largely speculative, yet the notion that swarms of tiny self-replicating robots could escape from laboratories and destroy our world comes up time and time again when nanotechnology is discussed with the public.

But researchers can take heart: the author who arguably did most to stir up the fears in the first place has publicly dismissed the scare stories. He says he wishes he had never coined the term 'grey goo', which is used to encapsulate them.

In his 1986 book *Engines of Creation*, Eric Drexler speculated that self-replicating robots with molecular-scale dimensions could be used to build everyday goods cheaply and efficiently. He also noted that such 'nanobots' could be dangerous if they escaped: "We cannot afford certain kinds of accidents."

The idea gained a lot of influence, and the term 'grey goo' appears regularly in popular accounts of nanotechnology. It was also the inspiration for *Prey*, a bestselling 2002 novel by Michael Crichton that is being made into a film.

But in a commentary published on 9 June (see C. Phoenix & E. Drexler *Nanotechnology* 15, 869–872; 2004), Drexler acknowledges that nanoscale manufacturing does not require self-replicating devices. "I wish I had never used the term 'grey goo'," he told *Nature*.

Drexler says that if he could write *Engines of Creation* again, he would barely mention self-replicating nanobots. Instead, he would focus on the possibility that nanoscale manufacturing will become a miniature version of conventional processes, in which a single device moves along a conveyor belt and is built up in a series of discrete operations.

The California-based author says he has gone public now because of worries about the way nanotechnology is being perceived. "Fears associated with that old scenario are interfering with current research," says Drexler. "Researchers resent it and I want to clean up the mess."

But Drexler's repentance may not change much. Experts who track public attitudes to nanotechnology say that some environmental groups have played on fears about 'grey goo' to publicize their more genuine concerns about who will benefit from the technology. "If the term wasn't here, the groups would come up with another phrase," says one expert. ■



Cost of living: spending US\$27 billion should prevent 28 million cases of AIDS by 2010.

Fighting AIDS is best use of money, says cost-benefit analysis

Munich Channelling money towards AIDS and malnutrition would be the most cost-effective way of tackling the world's problems, according to a group of eight prominent economists.

The two issues topped a ranking of aid priorities hashed out at a meeting in Copenhagen from 24 to 28 May. The panel estimates that it would cost US\$27 billion to prevent about 28 million HIV infections by 2010; but benefits from increases in the number of working adults and other factors would be almost 40 times as high.

Trade liberalization and malaria grabbed third and fourth place in the list; three schemes to tackle climate change, including the Kyoto Protocol, took the bottom spots.

The meeting (see *Nature* **428**, 110; 2004) was organised by Bjørn Lomborg, author of *The Skeptical Environmentalist*, a 2001 book that questions widely held beliefs about the environment. Lomborg, a critic of Kyoto, is head of the Danish Environmental Assessment Institute in Copenhagen.

♦ www.copenhagenconsensus.com

Boehlert happy with science despite elevation rumours

Washington The chairman of the US House Committee on Science, Sherwood Boehlert (Republican, New York), plans to keep his position even if tempted with a more prestigious chair, he says.

Boehlert, who has served 22 years in the House of Representatives, is next in line for the senior position of chair of the Permanent Select Committee on Intelligence. But Boehlert tells *Nature* that he will stick with science if re-elected. "I've made it very clear that my intention is to continue to chair the science committee," he says. "I'm in a very good position, and I enjoy it."

Boehlert is seen by many Washington lobbyists as an effective advocate for

science on Capitol Hill (see *Nature* **422**, 365; 2003). There has been speculation in recent months that he might soon leave for the intelligence committee.

Biological pathways database revamped

New York Scientists now have a single online database to help them discover how a given protein fits into the puzzle of human biology.

The Reactome database, which went live on 2 June, is a revamped, renamed version of the Genome Knowledgebase. The resource pulls together information on genes and proteins that work together in specific biological pathways, such as the process of copying a strand of DNA. The tool will help researchers to discover if a set of genes found to be active in a diseased tissue, for example, all contribute to the same biological process.

Reactome currently contains about 8% of all catalogued human proteins. Its coordinators — at Cold Spring Harbor Laboratory in New York and the European Bioinformatics Institute in Hinxton, UK — aim to expand this. "We're trying to cover all of human biology," says Lincoln Stein of Cold Spring Harbor.

♦ www.reactome.org

Corking result for French wine

Paris In a country where wine is drunk like water, the government is contemplating reclassifying the drink as a 'natural food' instead of alcohol.

French winemakers say sales of their national drink have been falling for years, thanks in part to a strict 1991 law on alcohol advertising and anti-drinking campaigns run by the government. They describe the situation as a 'crisis' and say that wine, as an important part of France's culture and history, ought to be given a helping hand.

Their complaints appear to have had an effect. Later this month, Prime Minister Jean-Pierre Raffarin will be presented with draft legislation proposing the reclassification. If passed, this will alleviate the heavy advertising restrictions.



US plans extensive cut to nuclear weapons stockpile

Washington The US government has announced plans for a significant cut in its stockpile of nuclear weapons between now and the end of 2012.

The plan is outlined in a report sent to Congress on 1 June by the National Nuclear Security Administration, which oversees the nation's nuclear stockpile. The report's details are classified, but Linton Brooks, the agency's administrator, says that the stockpile will be cut "almost in half". Currently, the United States is thought to have about 10,000 weapons.

The move has been welcomed by many. But Ivan Oelrich, who directs the Federation of American Scientist's Strategic Security Project, says it does not go far enough. "The US still has approximately ten times the number of warheads it needs," he says.

Sandia boss knew I was innocent, says worker

San Diego Sandia National Laboratories is being sued by an employee who claims she was made a scapegoat for security lapses, and that her career and reputation have been ruined in the process.

The New Mexico nuclear weapons lab

came under congressional scrutiny in early 2003 after whistle-blowers reported problems with security, such as napping guards and disappearing master keys. An external review committee claimed that Patricia Gingrich contributed to one such lapse — by helping to destroy evidence of an inappropriate romantic liaison. Gingrich says she was later demoted to a job with an \$11,000 drop in salary.

Gingrich's lawsuit, filed on 19 May, claims that lab president Paul Robinson knew the review's accusations to be false, but reprimanded her publicly in order to persuade Congress that steps were being taken to improve security. The lab has declined to comment on the case.

Correction

A News item in the 3 June issue of *Nature* (429, 490; 2004) erroneously states that Hanyang University Hospital's Institutional Review Board (IRB) was in violation of Korean Food and Drug Administration guidelines. The article says that the guidelines require IRBs to include more than one layperson. In fact they only require one, which this IRB had appointed. The error in the text was made by a Seoul-based translation company that *Nature* paid to translate the document. *Nature* asked the translator to verify the passage prior to publication, but the mistake was overlooked. We apologize for the error. ■

Frozen time

Researchers have pulled the oldest-yet core of ice from the Antarctic — giving us a 740,000-year record of the planet's climate. Gabrielle Walker braves the cold to find out how they did it, and what they hope to learn.

“Are you ready to go into the freezer?” It’s a warm summer’s day in Le Fontanil, southeastern France, but glaciologist Jerome Chappellaz is pulling on a hefty parka and snow boots. He punches a button and a door slides open into a vast commercial cold store, where whirring fans add a bitter wind-chill to the temperature of -25°C . Although we have come to look at ice, we’re first confronted with three floors of wooden pallets stacked up to the ceiling, bearing sides of beef, rings of yellow goat’s cheese and boxes of frozen raspberries.

“We work with a raw material that is very close to melting, so we’re always close to losing our samples,” says Chappellaz, from France’s leading glaciology laboratory — the Laboratoire de Glaciologie et Géophysique de l’Environnement (LGGE) in Grenoble, 15 kilometres from Le Fontanil. “But they have millions of euros of food in here, so there’s no way they will let it thaw.”

That’s just as well, because the samples Chappellaz is talking about are far more precious than cheese. Tucked away in the attic, in scores of cardboard cartons cross-wrapped with black straps like slightly battered Christmas presents, is one of the world’s most extraordinary archives — a record of ancient climate written into 15 kilometres of ice cores, taken from mountain glaciers and the frozen caps of Greenland and Antarctica. Chappellaz rummages in a carton, and lifts out a long thin wedge of ice, whose curved side is scored with spiral rings from each turn of the drill bit. Although it looks unremarkable, this core is setting the climate world buzzing. It comes from so deep in the Antarctic ice sheet that any bubbles have had the breath squashed out of them, and the ice is as clear as glass. Dating from roughly 740,000 years ago, it is also the oldest ice core ever retrieved.

The ice was drilled at Dome C in the East Antarctic ice sheet by a consortium called the European Project for Ice Coring in Antarctica (EPICA). It has nearly doubled the age record for an ice core — previously set by a core drilled 500 kilometres from Dome C, at the Russian station Vostok a decade ago. The flagship for ice-core research, the Vostok core extends back around 400,000 years, and contains four ice ages¹. It was a resounding success for researchers, demonstrating — among other things — that carbon dioxide levels in the atmosphere have marched in



exact step with temperature for hundreds of thousands of years².

Cold comfort

But, frustratingly, the Vostok core stopped in the middle of a period known as stage 11, which occurred about 425,000–395,000 years ago, and which the EPICA core has now fully embraced. Researchers are particularly keen to explore this period because it is the last time that the Earth’s orbit was very nearly identical to its present one, making it a potential model for our own climatic future. With its record of atmospheric gases frozen into the ice, the EPICA core could also reveal secrets of a mysterious climate transition that started a million years ago.

The site at Dome C, chosen for its exceptionally thick ice, is flat, white, and bitterly cold, with temperatures ranging from -50°C at the start of the season to -25°C in the middle of the Antarctic summer. But it is a surprisingly pleasant place to work. Located at the summit of a dome of ice, it escapes the chilling winds that sweep down

Antarctica’s slopes. There are heated tents for sleeping, a sauna, and unrestricted hot showers. “It’s really comfortable,” says chief driller Laurent Augustin. “I call it Club Med.”

The team’s drill contains a metal tube 3.5 metres long and 10 centimetres wide, attached to a mechanical drill bit. During each run, a few metres of ice are collected in the tube, and brought back to the surface, where the core is removed, labelled and stored.

Each stage of the drilling has its own problems. For the first 1,000 metres, air bubbles tend to form in the boundaries between ice crystals, making the core so brittle that the sudden release of pressure at the surface can make it explode and shatter.

Ice chips formed by the grinding drill can also clog up the mechanism, jamming it in place. This happened in the 1998–99 season, leaving the drill irretrievably stuck. The team was eventually forced to abandon the hole and start again, with a brand new drill head. More cautious after such a delay, they reduced the amount of ice pulled up to just three, two or even one metre with each run, making it a



Below zero: Jerome Chappellaz (right) and his colleagues start to analyse the ancient ice cores drilled at Dome C (above) from 1996 to 2003.



time-consuming process. When they had reached down more than 3 kilometres, it took more than an hour to bring the core up to the surface.

The team celebrated reaching 1,000, 2,000 and 3,000 metres with parties and champagne. And when field tests showed that the core had passed Vostok's age, there was a buzz with every metre that came up. "It was very exciting," says Eric Wolff from the British Antarctic Survey in Cambridge, UK, who spent two seasons as chief scientist at Dome C. "You knew you were getting stuff that had never been seen before."

But close to the end the project hit yet another snag. At 3,100 metres, the team realized that the ice was now close to melting point, thanks to geothermal heat. That's problematic, because melting ice can refreeze on the drill and trap the whole mechanism. They managed to retrieve some of the melting ice, but intend to return to the site next November for the remaining 100 metres, using a plastic bag filled with ethanol as antifreeze for the drill bit. "It's not very

elegant, but it works," says Augustin. That final section could reach back to 900,000 years or even further. "None of us knows what we'll find at the bottom of Dome C — that's why it's exciting," says Dominique Raynaud of the LGGE, who is the coordinator for the next stage of the project.

Climate clues

After the cores were drilled, they were sent to participating labs in Europe by ship, with a quarter-slice of the entire core remaining in Antarctica in case an accident befell the travelling ice. Some surprising results — including the core's extreme age — have already emerged³ (see also the article in *News and Views* on page 611), though much of the Dome C core remains to be studied.

The most eagerly awaited findings are from the samples of ancient air trapped in its ice — a record that is unique to ice cores. Chappellaz and his EPICA colleagues have been measuring everything they can in the ice's air, from carbon dioxide, methane and nitrous oxide to the isotopes of nitrogen and oxygen. However,

the measurements are laborious and time-consuming; so far, they have only managed to confirm Vostok's records in the top part of the core and reveal the composition of the atmosphere during stage 11.

That already shows that the amount of CO₂ in the air during stage 11 was similar to our own pre-industrial level³. As Earth's orbit was much the same then, this suggests that stage 11 was very like our present balmy interglacial, and that without the effects of global warming we would probably have to wait some 16,000 years for another ice age. "It's a wonderful window into the future as well as the past," says Raynaud.

Measurements of the ratio of isotopes of oxygen in the ice and the sizes of the ice crystals, both of which vary with temperature, have also shown that before 450,000 years ago the climate cycles seemed flatter, with a smaller difference between the extremes of ice ages and the warmer interglacials³. Researchers think this is a sign of the tail-end of a mysterious transition in our climate's history.

A million years ago, ice ages recurred about every 40,000 years — the same timescale on which wobbles in the Earth's tilt alter the amount of sunlight hitting its surface. But now our ice ages return every 100,000 years. This is the timescale on which Earth's orbit becomes more or less elliptical, but this does not provide enough of a change in the amount of sunlight to explain the ice ages. So why should the cycle length have changed?

One possible explanation is that CO₂ levels may have gradually fallen, making the world progressively colder, says Wolff. This would have caused successive ice ages to grow larger and larger ice sheets — something that marine records suggest has been occurring over the past 2 million years.

Eventually, the ice may simply have become too slow on its feet to respond to the 40,000-year beat, and could have been forced to adopt the more stately, 100,000-year driving force. The Dome C core should soon reveal whether this is right. "If carbon dioxide really is the key, this is the only convincing way of getting it," says Wolff.

EPICA researchers around Europe are now furiously crushing and measuring their ice to see what the greenhouse-gas record will reveal. "It's a strange feeling to look at a sample melting after we have extracted the air," says Chappellaz. "Water that fell as snow in the middle of Antarctica 700,000 years ago will end up in the Isère river in the middle of Grenoble. And every time we open one of our containers, a fragment of atmosphere that's hundreds of thousands of years old disappears back into the air."

Gabrielle Walker is a science writer based in London.

1. Petit, J. R. *et al.* *Nature* **387**, 359–360 (1997).
2. Petit, J. R. *et al.* *Nature* **399**, 429–436 (1999).
3. EPICA community members *Nature* **429**, 623–628 (2004).

Bioprospects less than golden

The search for potentially valuable natural products is flagging. Could new rules on ownership of these resources give it a boost? Rex Dalton investigates.

In the balmy atmosphere of the 1992 Earth Summit in Rio de Janeiro, a convention to conserve global biodiversity seemed like an idea whose time had come.

The Convention on Biological Diversity was agreed at Rio, in principle, as a framework that would help the world's biological resources to be utilized in a controlled and prudent way. Scientists, governments and commercial companies would work together in harmony, the convention's authors hoped.

But progress towards these goals has been painfully slow. Searching for genetic resources among plants and animals has never been easy: the richest biodiversity is found in exotic but sometimes lawless locales. Its study requires not just scientific perseverance but the construction of an intricate web of relationships with local people, landowners and government officials.

And the convention has done little to ease tensions that exist between scientists searching for potentially valuable compounds and officials in the developing countries where most bioprospecting takes place. If anything, this tension is growing as nations become more aware of their rights under the Rio convention.

Bioprospecting still takes place in "an environment of suspicion", explains Rohan Pethiyagoda, a biomedical engineer and conservationist who founded the Wildlife Heritage Trust of Sri Lanka. He adds that suspicion usually falls equally on corporate bioprospectors and academics — to the disadvantage of researchers such as taxonomists, whose main aims are non-commercial.

The mistrust is based on decades of "academic scientists bringing home suitcases full of leaves, mushrooms or whatever", says Daniel Janzen, an entomologist at the



Hide and seek: finding useful medicinal compounds in the Costa Rican rainforest has proven more difficult than anticipated.

University of Pennsylvania who conducts fieldwork in Costa Rica. The samples would sometimes be sold for cash, Janzen says.

The treaty has reined back such excesses. Most major corporations in the United States and in Europe officially adhere to its principles, even though the United States is not a party to the convention. But it has done little to quell poor nations' fears of exploitation. Nor has it yielded the anticipated bioprospecting bonanza. On the contrary, many pharmaceutical companies have withdrawn from the field for a variety of reasons, including doubts about its commercial benefits.

Getting it right

This has left the backers of the convention seeking new instruments to shore up its effectiveness. Developing nations, many scientists and even some companies that want to exploit biodiversity would welcome more specific rules on how the origins of valuable samples should be recorded and their benefits shared between the nation of origin, the discoverer and the commercial sponsor.

But such rules have powerful opponents, including government and pharmaceutical company officials in the United States, who claim that proposed rules would run counter to the existing US patent system. Their implementation — which was discussed earlier this year in Kuala Lumpur, Malaysia, at the seventh conference of the parties to the convention — is set to be a long, hard slog¹.

On the ground there are some successes, however. Just as the Rio convention was taking shape in 1992, Phyllis Coley, a plant ecologist at the University of Utah in Salt Lake City, launched a bioprospecting project in Panama to look for potential drugs against the parasites that cause malaria, leishmaniasis and Chagas' disease. Today, the project has evolved into a network of six laboratories, employing ten senior scientists, nearly 60 technicians and training dozens of local students.

Coley's team looks specifically at young plant leaves, which they have found to contain higher levels of compounds such as alkaloids². These bioactive compounds protect the plant against insect predators and can be potential drug candidates.

Support for the network comes primarily from the Fogarty International Center, a branch of the US National Institutes of Health that specializes in work overseas. From the start, Coley ensured that the project was deeply rooted in Panama. All its research is done there. Laboratories have been created, staff trained and more affordable assays developed to test plant samples for activity. Project managers are meticulous about transparency: each time a sample is to be transported out of the country, approval is sought from the Panama National Authority on the Environment.

Coley is proud of the project's track record. A fluorescence assay for antimalarial agents has been developed and patented, and



Take your pick: Panamanian biologists collect leaves in a project backed by the US National Institutes of Health.

is now being used by research groups in Africa³. The patent for the assay has been filed in the United States, with rights assigned to researchers and institutions in Panama. And the project team helped Panama to apply for World Heritage Site status to protect Coiba National Park, a region of islands and coral reefs on the Pacific Coast.

The training of young scientists has been the project's greatest achievement, says Luis Cubilla, a project scientist from the University of Panama. "That is very important for my

country," he says. "We are developing a core group that have the skills needed to build a research enterprise."

But in drug discovery — the project's central objective — results have been less stellar. After 12 years and US\$6.55 million investment, the project has yet to deliver a compound ready even for safety trials in humans. This lack of commercial prospects is fairly typical of otherwise successful projects worldwide.

A handful of promising agents have been identified, Coley says. A patent is being sought on a plant compound that shows activity against leishmaniasis⁴. Toxicity studies in mice are under way in Panama, and additional work will be conducted this summer at the London School of Hygiene and Tropical Medicine.

Coley remains optimistic that useful drug leads will eventually be found. In the meantime, the project is building up capacity on the ground. "Waiting for royalties is not going to help countries like Panama, because the timescale is too long," she says. "You have to train people, create jobs, and develop local awareness of biodiversity."

Protecting interests

The transparency of the Panama project and its depth of local involvement make it a useful model for others. In the Philippines, for example, bioprospecting has often been problematic. In 2000, for instance, three French scientists masquerading as eco-tourists were caught trying to smuggle out medicinal plants.

This context presents a challenge for a collaboration of Philippine and American scientists, who have a planning grant from the Fogarty centre to start bioprospecting in the Bataan National Park, which covers some 23,700 hectares of highlands 150 kilometres west of Manila. Research at the Bataan park is further complicated by concern that it might unfairly exploit the traditional knowledge of the indigenous Ayta people who live nearby.

Past experience in places such as the Philippines has made bioprospecting difficult for taxonomists, who are seeking to catalogue the world's plants and animals. Usually, their work is purely academic. But as nations learn to defend their genetic resources, local officials sometimes suspect them of biopiracy, and prohibit the export of specimens for analysis.

But Lourdes Cruz, a member of the Bataan project and a biochemist at the University of the Philippines, says considerable headway has been made. "We are the first to get prior informed consent from the Ayta people to work in the park," says Cruz.

A Bataan Technology Park has been established to house research and development facilities, and scientists from the

university's medical school and Marine Science Institute are engaged in project planning.

Nevertheless, says physician Michael Kron of Michigan State University in East Lansing, the principal US investigator on the project, it is likely to take all of the two-year period of the planning grant just to negotiate a collaborative research agreement with the Philippine Ministry of Health. "If we get an agreement, I will consider it a major accomplishment," says Kron, a specialist in infectious diseases.

All this preparation is necessary, bioprospectors believe, to avoid the pitfalls that await them in countries that are only now learning to assert their rights under the biodiversity convention (see "Bermuda gets tough over resource collecting", overleaf).

One reason for the regulations that may be added to the convention is to secure the conservation of biodiversity. But another is to create a more robust framework that will entice major drug manufacturers back into bioprospecting.

Few large pharmaceutical companies now have serious investment in the search for natural products. Coley, for example, had spent years negotiating with Monsanto, of St Louis, Missouri, on a legal agreement for any products discovered in Panama, only to see the deal collapse when the company's natural products section closed. "Suddenly, we were

frantically looking for a corporate partner," she recalls. Swiss-based pharmaceutical company Novartis eventually stepped in, agreeing to contribute \$50,000 for laboratory equipment. She says she is unsure why companies are reluctant to participate. "It's strange," she shrugs. "It's almost a free lunch for them."

In 1991, amid much fanfare, pharmaceutical giant Merck, based in Whitehouse Station, New Jersey, made a deal with Costa Rica to provide about \$1.2 million over ten years for bioprospecting and conservation. But no comparable arrangements have been announced since. And most companies have taken a more tentative approach. Some, including Monsanto and New York-based Bristol Myers Squibb have shut down their natural products divisions entirely.

Merck itself has now halted investment in the Costa Rica project, providing its last grant of \$130,000 in 2001. A company spokeswoman said that no products had come out of the project. Officials at Merck and the other two firms refused to discuss in detail why they have folded their bioprospecting tents.

But Lila Feisee, director of intellectual property at the Biotechnology Industry Organization in Washington DC, says that she suspects the push for benefit-sharing by developing nations has given drug companies "a negative image of bioprospecting".

"Companies need incentives," says Feisee.

"Waiting for royalties is not going to help. The timescale is too long. You have to train people, create jobs, and develop local awareness of biodiversity."

— Phyllis Coley

"But those incentives are being undermined. When risks outweigh the benefits, companies will do something else."

Natural remedy

This choice frustrates scientists who believe that natural products remain the most promising source of new drug treatments in the long term. David Newman, a chemist and leading authority on natural products who runs the marine and microbe collections of the National Cancer Institute in Bethesda, Maryland, blames the drug company pull-outs on corporate inertia. "There is a fear of gambling, then failing," he says.

In a study⁵ published last year, Newman and his co-authors wrote: "The decision on the part of several pharma companies to get out of the natural products business is gross foolishness. The utility of natural products as sources of novel structures, but not necessarily the final drug entity, is still alive and well."

Reviewing drug discovery over the 22 years up to 2002, Newman's study found that almost two-thirds of anticancer agents being investigated as drug candidates were derived from natural products.

One reason for the low level of interest of large companies, bioprospecting advocates believe, is the lack of a firm framework for benefit-sharing between host nations, scientists and commercial companies.

At this year's Kuala Lumpur conference,



Local involvement: Lourdes Cruz (right) explains the Bataan project to Aya tribespeople in the Philippines.

representatives of 188 nations agreed to try to build such a framework, to be considered at the next meeting in Brazil in 2006. Matthew Jebb, acting director of Ireland's National Botanic Garden in Dublin, who led the European Union delegation in the negotiations, says that "a vital underpinning can be created to dispel that air of suspicion" through such a benefit-sharing agreement.

A clear title document for each compound discovered would be an important element of such an agreement. Advocates envision a

document that would follow compounds around like a passport, stating where they came from and who holds rights on them.

Bioprospectors say that this arrangement could help entice drug companies back into the game. At the Kuala Lumpur meeting, the United States resisted the concept of a new, benefit-sharing agreement under the convention. But Leonard Hirsch, a Smithsonian Institution economist who is vice-chair of the US delegation, says he remains open to the idea. The United States is "seriously engaged in an analysis of the most efficient and cost-effective mechanism for a certificate of origin", he says. At upcoming meetings in Thailand next February and in 2006 in Spain, Hirsch says, the United States "looks forward to honing proposals and developing a user-friendly certificate system".

But despite such reassuring words, concern remains about the future of the convention's goal. "My biggest fear is that no new access and benefit-sharing agreement will be reached," says Jebb. "There is a great urgency — a finite biological resource is disappearing. It is going to cost all of us if no regime is enacted."

Rex Dalton is Nature's US West Coast correspondent.

1. Dalton, R. *Nature* **427**, 769 (2004).
2. Coley, P. et al. *Front. Ecol. Environ.* **1**, 421–428 (2003).
3. Corbett, Y. et al. *Am. J. Trop. Med. Hyg.* **70**, 19–124 (2004).
4. Montenegro, H. et al. *Planta Med.* **69**, 677–679 (2003).
5. Newman, D. J., Cragg, G. M. & Snader, K. M. *J. Nat. Prod.* **66**, 1022–1037 (2003).
6. Venter, J. C. et al. *Science* **304**, 66 (2004).

Bermuda gets tough over resource collecting

When Craig Venter steered the *Sorcerer II* into Bermudian waters early last year, he was searching for ocean microbes. And he found plenty: in a paper published in March, he recorded 1,800 microbial species, including 148 previously unknown bacteria⁶.

But his voyage into the Sargasso Sea also took the genomics pioneer into uncharted waters. The rules on bioprospecting in this small British protectorate are still a

work in progress. And experience with expeditions such as Venter's has prompted Bermuda to temporarily shut down some research projects until it strengthens its regulations.

Bermuda is now rewriting its scientific collection rules completely, in preparation for joining the Convention on Biological Diversity as a protectorate of the United Kingdom, which has already adopted the convention. And lessons learned from Venter's scientific expedition and from a separate commercial project started in 1999 by Diversa, a San Diego firm seeking drugs from microbes, technology tools and industrial chemicals, will influence the



Craig Venter (left) and Anthony Knap aboard *Sorcerer II*.

formation of these rules.

Both Venter's group and Diversa gathered marine samples under the umbrella of a long-standing collection permit held by the Bermuda Biological Station for Research in St. George. Nearly 100 years old, the station gets most of its funding from NASA and the US National Science Foundation.

In 1999, Diversa struck a deal with the station to bioprospect in the waters off Bermuda.

But Jack Ward, the official at the Bermuda Ministry of the Environment responsible for developing bioprospecting policy, says that the government only learned of Diversa's project from a newspaper article published at the time. After enquiries, Ward says, station officials explained that Diversa was only studying organisms that could be found in many locations in the Atlantic Ocean, not looking for new ones. Subsequently, Bermuda didn't require Diversa to secure a government permit.

But in May this year, Ward learned that Diversa was marketing a biotechnology tool called DiscoveryPoint Fluorescent Proteins, which was based on a protein collected from

a coral in Bermuda, and for which the firm is seeking a patent. The research station is to get a 1% royalty, but the government and people of Bermuda will get nothing.

Meanwhile, Venter sailed into the Sargasso Sea in February 2003. The organisms collected were shipped to Maryland, where a US Department of Energy grant paid to have them sequenced at the Venter Science Foundation Joint Technology Center.

"We do this as part of discovery to enhance science," says Venter. "We are trying to benefit every country we work in." Bermudian officials don't fault Venter, who says that the Bermuda Biological Station's director, oceanographer Anthony Knap, assured him that no government permits were required — advice that Knap confirms. The microbial DNA sequences are being placed in the publicly accessible GenBank for scientific use.

But Ward says that the ministry is unhappy about the outcomes of both projects, and that Bermuda is revoking the station's collection permit. A new permit, with stricter controls, will be in place within the next month.

Bermudian officials regret what they regard as a lost opportunity. "There is a value issue here," says Ward. "Something that held value has been put in the public domain and made valueless for the people of Bermuda."

Small countries receive even less of a fair deal

Governments and companies should intervene to control the price of lab supplies.

Sir—We welcome your Editorial “A fair deal for all” accompanying the survey reported in “High prices of supplies drain cash from poorer nations’ labs” (*Nature* 428, 451 and 453; 2004).

These address a topic of great importance to scientific researchers in less developed countries. Recognizing this problem is the first step towards finding and implementing some solutions.

Your survey compared the costs of laboratory equipment and materials in four countries — Germany, Poland, the United States and Brazil — and revealed some differences between the poorer and wealthier nations. Poland is, however, the largest market in Eastern Europe; in smaller markets, such as Croatia, the situation can be even worse for scientists.

When prices of laboratory equipment and materials in Croatia are compared with those in wealthier states, the differences turn out to be still more

pronounced. There are, for sure, a few items that are exceptions to this general rule, but in our experience most lab equipment and materials show similar, if not bigger, price differences.

For example: in Croatia, 500 g albumin, bovine fraction V protease-free, costs approximately €1,458 (US\$1,796) compared with the manufacturer’s price of €992 (using an exchange rate of 7.5 kuna to the euro and excluding value-added tax). Similarly, 0.5 ml mouse monoclonal anti-fish CYP1A peptide costs about €1,112 in Croatia, compared with the producer’s price of €480.

These represent the best prices that we have found for these supplies.

One should also consider the low levels of research funding in poorer countries. In Croatia, the government’s expenditure on research grants is many times lower than in more affluent states.

For this reason, it is no exaggeration to say that the only way a research group from

one of Europe’s poorer nations can publish its work in prestigious journals is to collaborate with Western researchers, who benefit from higher funding and lower costs of materials.

In Croatia, the prices of laboratory equipment and materials could be controlled by the state to prevent local distributors selling products at inflated prices. In addition, the producers themselves could intervene and encourage local distributors to sell at the manufacturer’s price.

This would help eastern European researchers get better value for their money, making the scientific game a little more just.

Stipan Jonjić*, Luka Travenč†

*Department of Histology and Embryology, Faculty of Medicine, University of Rijeka, Brace Branchetta 20a, 51000 Rijeka, Croatia

†Department of Environmental Health, Faculty of Medicine, University of Rijeka, Brace Branchetta 20a, 51000 Rijeka, Croatia

Fair deal: local factors add to price of supplies

Sir—In response to your News story “High prices of supplies drain cash from poorer nations’ labs” (*Nature* 428, 453; 2004), I would like to make clear that Amersham Biosciences (now GE Healthcare) maintains a transparent and consistent pricing policy as part of its ongoing commitment to the global scientific community and customer service.

As with any international business, many factors can cause differences in the final price paid, such as currency valuations, quantities, import duties, local taxes, distribution channels and the local competitive situation.

The list prices of the two Amersham Biosciences products that were highlighted in this News story — the EPS 601 power supply and the Ultrospec 3300 pro spectrometer — are actually consistent across those markets (allowing for currency fluctuations).

Below are the list prices we publish in the relevant currency for the four countries quoted in your News story.

EPS 601:

United States and Brazil US\$963;
Germany and Poland €860.

Ultrospec 3300:

United States \$12,187; Brazil US\$11,812;
Germany and Poland €8,680.

Our list prices are publicly available in our catalogue and online, and are broken down by region. Because of fluctuations in currency valuation, these should not be represented in any one currency for different regions.

Andrew Carr

GE Healthcare, 800 Centennial Avenue,
Piscataway, New Jersey 08855, USA

Fair deal foiled by taxes and exchange rates

Sir—We would like to provide some information to contest the statements about Eppendorf products made in your recent News story about the price of laboratory supplies in poorer countries (*Nature* 428, 453; 2004).

The Brazilian price stated for a centrifuge 5415D (US\$3,110) is not the regular price for this item. The centrifuge 5415D is sold in local markets by our official dealers on average for \$2,500. Some research institutes are able to import products directly in order to avoid taxes. In this case, the centrifuge would cost \$2,100.

It should be noted that Brazil has a very high tax burden (14% import taxes, 8% manufactured taxes, 18% sales tax). These taxes occur in cascade and must be paid when the dealer receives the product (each tax is added on top of the previous tax).

Depending on the product, these taxes amount to 50–60% of the final product price. There is currently no legislation to assist researchers buying scientific products, although clinical labs profit from a law granting certain tax benefits.

As Eppendorf is a European company, there is always some variation in price due to the exchange rates. Dealers buy products from Eppendorf in euros, whereas Brazilians are more used to US dollars or their own currency, the real (R\$). So we should take into consideration a series of exchange rates: R\$3.79 to the euro, R\$3.17 to the dollar and US\$1.20 to the euro. At present, this situation is extremely unfavourable for us.

Finally, there are non-authorized companies selling Eppendorf products in Brazil, whose prices we cannot control.

Josely Chiarella

Eppendorf do Brasil Ltda.,
Praça Panamericana no. 42, Conj. 13,
05461.000 Alto de Pinheiros,
São Paulo-SP, Brazil

The aim of our news coverage was to discover what prices researchers in different countries actually have to pay for a given piece of equipment. So we asked the researchers in each country to contact their authorized local retailers, just as they would under normal circumstances. The prices quoted in our story therefore reflect the authentic quotes they received — Editor, *Nature*.

Isolation is not the answer

International scientific collaboration is the best defence against bioterror.

Thomas May

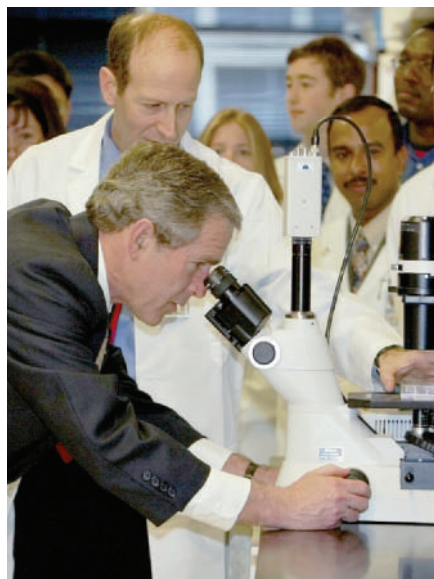
The fear of bioterrorism is increasing scientific isolationism in the United States. New restrictions on the publication of sensitive information relevant to biological weapons, on access to 'select' biological agents for research, and on the training of scientists from specified countries are some examples. Although restrictions on scientific activities might make sense in the context of nuclear-weapons proliferation, they may end up being counter-productive for the United States' defence against bioterror.

Biological terrorism poses a unique threat in that the devastation caused by the release of a biological agent is unlikely to be confined to the event itself, but will depend on the ease with which the disease spreads. Infectious disease cannot easily be restricted to any location, region or even nation. The World Health Organization (WHO) has documented numerous cases of global disease spread, including that of severe acute respiratory syndrome (SARS) in 2003. This virus disease originated in southern China and spread to nearly 30 countries, resulting in 8,098 infections and 774 deaths over nine months¹.

A bioterror attack would probably be much worse than this natural epidemic, as the infectious agents used and the manner of their release would be designed for maximum effect. If terrorists released a biological agent in a region where quick identification of a disease outbreak was unlikely, they could exploit the ease of international travel to spread the disease to 'targeted' countries once it had established a sufficiently strong foothold to make containment difficult. Consequently, attention to the global dimensions of bioterror threats is particularly important, including strengthening international means to identify and contain outbreaks of infectious disease.

To date, national-security experts have considered the risk of this sort of attack to be fairly low². It was thought that its indiscriminate nature — placing non-target populations, including the terrorists themselves, in danger — would undermine popular support for the terrorists' agenda. This analysis, however, reflects old expectations of terrorist behaviour based on rational self-interest — rules that simply do not apply to modern terrorists.

Although many terrorists try to limit casualties for pragmatic reasons, emerging terrorist groups often have radicalized agendas that pay less, if any, concern to public support. In addition, today's terrorists have consistently demonstrated a willingness to die (and to



Do President George W. Bush's policies for combating bioterror have the right focus?

kill innocent civilians) to achieve their goals.

In this context, is the United States developing the best strategies for preventing bioterrorism, and for responding to and containing bioterror attacks? Homeland defence priorities³ have emphasized diplomatic, intelligence, law-enforcement (including disruption of terrorists' financial networks) and border-control measures designed to keep potential biological weapons out of the hands of terrorists. These are valid efforts but they do not fully address the international dimensions of modern bioterrorism and the most likely route by which an attack will reach the United States. It is also widely accepted that the 'open' nature of US society creates vulnerabilities to terrorism³. But when preparing ourselves for a bioterror attack, an open academic and educational system is one of our most important defensive strengths. New ways of thinking about security are sorely needed.

Recognition of the true international nature of the bioterror threat should make the United States take a leading role in training foreign scientists, medical professionals and public-health personnel to build a global capacity for identifying and containing disease outbreaks. This must occur at several levels. First, the US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, must be better equipped to provide significant support and training to international public-health personnel. Although such support has long been a focus of the CDC, the agency does not have the resources to expand these efforts: in 2003 the Council

on Foreign Relations⁴ reported that even domestic training is "drastically underfunded". Second, efforts at the WHO to improve infectious-disease surveillance and containment must become priorities for the United States. The unique features of bioterrorism make practical improvements to international healthcare of equal strategic importance to traditional diplomacy.

Perhaps most important, we must take care to protect the open nature of our academic systems, and to avoid placing undue barriers on the training and education of foreign scientists and medical personnel. Although some restriction is necessary, attempts to control scientific expertise must be balanced with the need to promote security through scientific progress, such as the development of new tests and treatments to identify and contain disease outbreaks.

Some may argue that this will grant terrorists access to sensitive information and expertise, but it does not increase such risks significantly: even with restrictions it is relatively easy to find individuals willing to pursue biological weapons research for the right price. Ken Alibek, a former leading Soviet biowarfare scientist, reported⁵ the recruitment of former colleagues by several countries, including Iran and North Korea. In reality, we cannot control access to biological weapons expertise through control of domestic science alone.

In contrast, the dependence of US science on foreign scientists is such that biodefence research will be inhibited if we continue down a road of scientific isolationism. Apart from the obvious barriers that restrictions on access to scientific information and tools place on research, restrictions on scientific training for foreign nationals will delay those countries from developing expertise crucial to identifying and containing disease outbreaks — key to any global strategy against bioterrorism. What is required is the proliferation of scientific training worldwide, not scientific isolationism. ■

Thomas May is at the Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA.

1. World Health Organization http://www.who.int/csr/sars/country/table2003_09_23/en
2. Clarke, R. *Against All Enemies: Inside America's War on Terror*, Ch. 7 (Free Press, New York, 2004).
3. US Department of State. *National Strategy for Combating Terrorism* (February 2003). <http://usinfo.state.gov/topical/pol/terror/strategy/#intro>
4. Rudman, W. et al. *Emergency Responders: Drastically Underfunded, Dangerously Unprepared*. Report of an Independent Task Force Sponsored by the Council on Foreign Relations (2003).
5. Alibek, K. & Handelman, S. *Biohazard*, Ch. 20 (Hutchinson, London, 1999).

Current affairs

The rise to fame of the El Niño climate phenomenon.

Our Affair with El Niño: How We Transformed an Enchanting Peruvian Current into a Global Climate Hazard

by S. George Philander

Princeton University Press: 2004. 296 pp.

\$26.95, £17.95

Michael J. McPhaden

The 1997–98 El Niño was a watershed event in the history of climate research. It was by some measures the strongest El Niño on record and by far the most publicized. El Niño is a periodic disruption of the ocean–atmosphere system in the tropical Pacific that affects weather and climate around the globe. Once a scientific curiosity studied by a handful of specialists, by 1998 El Niño had become a household word recognized by people from every corner of the planet.

What accounts for the extraordinary public interest in El Niño? The answer lies in the scientific advances of the preceding two decades. The 1982–83 El Niño, which was the strongest of the century until then, was neither predicted nor even detected until nearly at its peak. This failure stunned the scientific community, which at the time was planning a major ten-year international research programme to study El Niño. The Tropical Ocean Global Atmosphere (TOGA) programme, when it was finally launched in 1985, placed a premium not only on developing a deeper understanding of El Niño, but also on implementing new observational and seasonal prediction capabilities. TOGA was largely a success. The observing system it engendered captured the explosive growth and epic magnitude of the 1997–98 El Niño in real time and, once under way, computer models were able to forecast the subsequent development of the El Niño two to three seasons in advance.

George Philander, author of *Our Affair With El Niño*, was part of the community of oceanographers and meteorologists who helped to shape this exciting period of rapid scientific progress. He writes with the enthusiasm of an eye-witness and the authority of an expert. The book skilfully weaves together descriptions of El Niño physics, the historical backdrop that led to widespread interest in El Niño, and philosophical perspectives on the role of scientific research in addressing present-day environmental problems.

A central purpose of the book is to describe simply but accurately what El Niño is, how it works, and what its consequences are. El Niño's hallmark attributes are unusually weak trade winds and warm sea surface temperatures in the tropical Pacific. It is the



The El Niño effect: temperature changes in the Pacific can trigger severe droughts in Vietnam.

warm phase of what is often referred to as the El Niño/Southern Oscillation (ENSO) cycle, which results from dynamical feedbacks between the upper ocean and the overlying atmosphere. La Niña, the opposite phase of the ENSO cycle, is associated with stronger than normal trade winds and unusually cold sea surface temperatures in the tropical Pacific.

Deciphering the underlying physics of the ENSO cycle is an ongoing challenge for the research community, and explaining it coherently to non-specialists is even more difficult. Philander uses easily understandable analogies, such as that of a swinging pendulum, to highlight key aspects of the oscillation. He also draws on metaphors from literature, music and painting to illustrate some of the physical concepts and research methods used in the study of El Niño.

As Philander observes, however, our fascination with El Niño derives not only from the intricate interplay of oceanic and atmospheric dynamics, but also from its effects on patterns of weather variability around the world. El Niño changes the spatial distribution of tropical Pacific rainfall, the effects of which are felt directly in the tropics and indirectly at higher latitudes through atmospheric 'teleconnections'. It is through these impacts that El Niño science intersects with human affairs.

Disparate parts of the globe may experience droughts, floods, heatwaves and intense storms that can be causally related to El Niño during periods of unusual warmth in the tropical Pacific. Likewise, El Niño-related weather patterns can create environmental conditions favourable for wildfires, outbreaks of infectious diseases and the degradation of air and water quality. The many

faces of El Niño that were highlighted in the press and on television in 1997–98 helped to inform the public about the phenomenon and its global consequences.

As the 1997–98 El Niño began to unfold, the ready availability of long-range forecasts prompted many individuals, businesses and governmental organizations to take preventive measures in anticipation of El Niño's impending onslaught. Unfortunately, there are many factors besides El Niño that govern the climate system. Also, El Niño forecasts are by their nature probabilistic, a concept that is difficult for forecast-providers to convey and for forecast-users to understand. Our climate crystal ball is, at best, still fuzzy, so there are inherent risks involved in using these forecasts to guide important decisions.

Philander offers a sceptic's view of the El Niño prediction enterprise, examining at length the failed 1997–98 drought forecast in Zimbabwe and its consequences, but not giving equal time to forecast successes elsewhere. The reader may therefore wish to refer to Stanley Changnon's book *El Niño 1997–1998* (Oxford University Press, 2000), which attempts to quantify forecast benefits in a case study focused on the United States.

Also, Philander portrays El Niño as predominantly perilous, hence the book's subtitle, *How We Transformed an Enchanting Peruvian Current into a Global Climate Hazard*. This characterization is fair but incomplete, ignoring the many opportunities that El Niño creates. The author refers to the positive aspects mostly in the nostalgic past tense, citing the *años de abundancia* (years of abundance) resulting from rains in normally arid regions of Peru, as described by nineteenth-century writers. Yet for the United States, the 1997–98 El Niño resulted

in far more economic gains than losses, and fewer fatalities, than during other years because of the reduced number and intensity of landfalling Atlantic hurricanes and the record winter warmth in the Midwest.

Perhaps the book's most important message is that: "The solutions to serious environmental problems will elude us unless we are all aware of, and respect, the profound differences between the world of science and human affairs." This harks back to

C. P. Snow's lectures on the 'two cultures' but with a twist: whereas Snow viewed science and technology as a panacea for solving the world's great social problems, Philander recognizes that science and technology are only part of the solution. Effective use of scientific information to benefit society must also reckon with prevailing cultural values and political imperatives.

Our Affair with El Niño is a very readable, entertaining and instructive book that will

appeal to scientists and non-scientists alike. The author does not shy away from controversy in expressing his opinions about the sociological and political aspects of climate research. Whether or not you share his opinions, Philander unquestionably excels at describing the physics of the ocean, the atmosphere and El Niño in lucid terms. ■

Michael J. McPhaden is at the NOAA/Pacific Marine Environmental Laboratory, Seattle, Washington 98115, USA.

A helping hand

Hearing Gesture: How our Hands Help us Think

by Susan Goldin-Meadow

Belknap Press: 2003. 304 pp. \$29.95, £19.95, €27.70

Eve Sweetser

Over the past two decades, researchers have produced overwhelming evidence that the gestures we use as we speak are integrally connected to both our speech and our thought processes. Susan Goldin-Meadow has been at the forefront of this new scientific direction. In *Hearing Gesture*, she provides a synthesis of her decades of work on gesture studies. It is a welcome scholarly arrival for gesture researchers, and should be important news to social and cognitive scientists, who so far have paid little attention to the gestures that accompany speech.

Hearing Gesture is an engaging (even suspenseful) read and, with its clear and informal style, should be largely accessible to non-experts. It centres around four primary questions. First, is gesture really a window on thought? If it is, do most people (as opposed to just researchers) read gesture? Does gesture also help the speaker's own cognitive processes — and if so, how? And finally, what are the differences between the gestures that accompany speech and visual gestures used on their own? Goldin-Meadow examines these questions — her answer to the first three is 'yes', by the way — in the lab and in everyday settings such as the classroom. In so doing, she looks at the communication of infants, children and adults, including sighted and blind, deaf and hearing, and normal and cognitively impaired subjects.

Goldin-Meadow has pioneered ways of studying gesture, one of her signature methods being the comparison of 'matched' gestures, which overlap in meaning with the accompanying speech, and 'mismatched' gestures, which either complement or conflict with the linguistic meaning. With Breckie Church she observed children explaining their answers to piagetian conservation tasks (conservation of mass, number or volume when physical appearance is altered). Some children produce mismatched gesture—



Handy hints: gestures provide information that can make it easier to understand speech — or even replace it entirely.

speech pairings. For example, they say that a tall, thin container has a large volume "because it's taller", but simultaneously make a gesture indicating width; this shows awareness that the container's width, as well as its height, is relevant to the quantity of water it holds. These children, it turns out, are the ones who are most ready to learn about conservation, either by instruction or experimentation (*Cognition* 23, 43–71; 1987).

The contrast between matches and mismatches turns out to be a remarkable tool. Goldin-Meadow's later studies show that matched gestures lower the cognitive load on the speaker and speed the listener's comprehension, whereas mismatched gestures raise the load on both sides of communication, which makes sense because they bring in another cognitive model besides that presented in speech. However, Goldin-Meadow argues that mismatches are advantageous in other ways. Because hearers do 'read' gestures and process the information expressed (as also shown in earlier work by David McNeill and Adam Kendon), mismatched gestures not only allow speakers to express models that are inaccessible to speech but also give listeners access to those models, with the added advantage of providing potential feedback to speakers.

The use of the term 'mismatch' presents difficulties from time to time — it is regrettably not always clear which kind (comple-

mentary or conflicting) is most relevant in a given study. The author remarks that gestures rarely correspond precisely with words in meaning. Taken to its logical conclusion, this should mean that complementary mismatches, like matches but not like conflicting mismatches, show overlap between gestural and linguistic meaning — the boundary between matches and mismatches is perhaps presented as more tidy than it really is. However, there is careful differentiation in some crucial cases, such as the examination of cognitive load, and Goldin-Meadow comments that apparently conflicting mismatches often reflect different aspects of a potentially unified larger cognitive framework.

Another strand of Goldin-Meadow's work has been the examination of purely gestural communication, including that of deaf children with hearing parents. She compares their individual gestural systems with conventional signed languages and with hearing gesture that has taken over the communicative load. This provides rich evidence from several domains for McNeill's claims that gesture becomes 'language-like' when it takes on the primary informational load of communication. Gesture becomes conventionalized, segmented and even 'grammaticized' — the gestural systems of orally raised deaf children have a basic grammatical structure.

I have touched on only a few of Goldin-

Meadow's projects and methodologies. Readers will be impressed by her extraordinary combination of thoughtful insight, experimental ingenuity and immense persistence and dedication to the search for knowledge. She and her co-workers are currently researching such applied issues as the need to interpret children's gestures alongside speech in legal and psychiatric questioning, and also the degree to which adult questioners' gestures influence children's output.

Fully recognizing the vast unknown areas awaiting gesture researchers' attention, Goldin-Meadow presents multiple viewpoints where there is disagreement, in the examination of signed language gesture, for example, or the connection between gesture and lexical access. And she continues to push the boundaries of her field, raising new questions alongside the ones she answers.

Hearing Gesture stands beside McNeill's *Hand and Mind* (University of Chicago Press, 1992) and *Language and Gesture* (Cambridge University Press, 2000) as a milestone in the study of gesture's relationship with language and thought. It may help to reshape the basic premises and methods of psychologists, linguists and other social scientists. ■

Eve Sweetser is in the Department of Linguistics, University of California, Berkeley, California 94720-2650, USA.

Bedside stories

The Body in the Library: A Literary History of Modern Medicine

edited by Iain Bamforth
Verso: 2003. 418 pp. £16, \$25

John Carmody

Ever since C. P. Snow's *ex cathedra* declaration of the constraining existence of two cultures — the unbridgeable separation of the sciences and the humanities — scientists have been forced on to the defensive. It rarely seems to matter greatly that 'liberal' and 'literate' people know little of science. And it rarely seems to be appreciated how creative and profound science can be — or how torpid and turgid many humanities texts are.

Medicine should bestride both of Snow's cultures, binding them with imagination and the truest humanity. At its best it can do this: in the divine hierarchy of classical Greece, Apollo was a god of both medicine and music. Iain Bamforth is a medical doctor as well as an essayist and poet, and his new anthology, *The Body in the Library* — despite a title that suggests crime and malfeasance — is a superb reminder that the creativity of physicians flows beautifully beyond the consulting room or laboratory. Not all of the authors in his "history of medicine as told through literature" are doctors or medical scientists,

Exhibition

Living with luminescence

Frazzled by the frenetic lifestyle of the future? No problem, a luxury lounge bathed in soft, calming bioluminescence will soothe your worries away. That's the art concept created by Sydney-based microbiologist Kathy Takayama and artist John Nicholson, who unveiled their exhibit — entitled *LuxCorp* — at the Canberra Contemporary Art Space on 14 May.

With their prototype light-emitting furniture (right), they present a vision of a future in which human-bacterial interactions are taken to a new dimension. The light emitted is a simulation of the bioluminescent by-product of bacterial communication. The exhibit is part of *Metis*, the



Australian festival of art meets science, and runs until 18 June.
Carina Dennis

but I found the pieces that were written from the inside to be telling and touching.

Bamforth has confined his selection to material that was "produced after the event which turned medicine into a public utility — the French Revolution". Although the "modern medicine" mentioned comes from the nineteenth and twentieth centuries, Bamforth does seem to share my view that modern medicine really began in 1543 with the publication of the great anatomical text *De humani corporis fabrica* by Andreas Vesalius. More important, though, is the scope of Bamforth's trawl, which has gathered together some lovely work that would normally escape the attention of anglophone readers (or any who are monolingual). If I note a lacuna, it is a lack of direct, rather than narrative, writing about medical science: surely a Nobel prize address or two would have been worthy of our attention?

After an exhilarating and provocative introduction, and a debatable beginning with Dickens, this collection plunges us into a banquet of many delights. One of the earliest (from 1812) is an urgently written letter by Fanny Burney, recounting a dreadful operation. This is matched by the wonderfully ironic *The Cure* from 1810 by Johann Peter Hebel, and we are then brought abruptly to earth by the episode of the botched Talipes (club-foot) operation in *Madame Bovary*. The worldly wise Lytton Strachey shows us Florence Nightingale in her great days in the Crimea: it is chastening to be reminded of the old (but enduring) strife between doctors and nurses. Léon Daudet reverses that mirror to show us the *mélange* of the magnificent and the malicious in the nineteenth-century neurologist Jean Martin Charcot, putting me edgily in mind of one of my own pedantic and francophile clinical professors.

Not everything here is a success, though. There is a wooden-headed piece by G. K. Chesterton, a rather pointless letter by Anton Chekhov, some self-importance (*Illness*) from Virginia Woolf and vacuity from Alain, the pseudonym of philosopher Emile-Auguste

Chartier. Set against those stumbles are a surreal story by Franz Kafka, its tone utterly unexpected from its bland title, *A Country Doctor*; a chilling medical-social narrative commentary by William Carlos Williams (*Jean Beicke*), the droll *My Double* by Alfred Döblin and the dazzlingly cynical *Oedipus in Danger* by Robert Musil.

In *Irrationalism and Modern Medicine*, Gottfried Benn coolly and prophetically asks about the point of extending an unexamined and spiritless life. This, surely, is a daily question for contemporary medicine, but when we do die — no matter how long we delay that fatal event — is it important to do it "well"? Can we achieve that, and does it matter? After reading George Orwell's *How the Poor Die* we have to wonder about the manner of our death as a reflection or an inevitability of our life. Like Dezső Kosztolányi's *The Stranger*, Orwell's dispiriting story should refocus the physician's mind on to the challenge of thinking of the patient as a brother. For all his empathic fame, R. D. Laing's *Clinical Vignettes* made little impression on me and certainly did not achieve a comparable insight, in marked contrast to the honest human bewilderment of the extracts from Miguel Torga's *Diary*.

The episode that I treasure most in this inspiring and yet unsettling anthology is *Heart Suture* by Ernst Weiss. With the deftness of an accomplished surgeon, it whips us from the set-piece detachment of classical 'grand rounds' to an emotional wrench when a young anaesthetist realizes that the God-Professor's emergency patient is his former lover. The professional and the personal are balanced with admirable finesse, with a cogent resonance for the reader of bitter experience or the relief of a "There but for the grace of God..."

I am left, then, with an aphorism (which Bamforth quotes): "Perhaps one day we will realize there was no art but only Medicine." ■
John Carmody is in the Faculty of Medicine, University of New South Wales, Sydney, New South Wales 2052, Australia.

Causing a commotion

Niche construction: do the changes that organisms make to their habitats transform evolution and influence natural selection?

Kevin N. Laland, John Odling-Smee and Marcus W. Feldman

Evolutionary biology has always generated vigorous debate, from the arguments about Lamarckian inheritance during the 1900s, to the more recent disputes over punctuated equilibria and group selection. The latest spat — a storm in a teacup compared with these earlier controversies — concerns the seemingly prosaic observation that the activities of organisms bring about changes in their environments — a process known as niche construction. In January this year, ecologist John Vandermeer described the niche-construction perspective as “a major breakthrough”, yet an earlier review by Laurent Keller in *Nature* (425, 769–770; 2003) dismissed it as hyperbole, and evolutionary biologist Richard Dawkins warns of its “pernicious” reasoning in a forthcoming article. It is a subject that typically provokes strong and polarized responses. What is the fuss about?

At the heart of the controversy lies the nature of causality in evolution. Adaptation is conventionally seen as a process by which natural selection shapes organisms to fit pre-established environmental ‘templates’. The causal arrow points in one direction only: it is environments, the source of selection, that determine the features of living creatures.

Yet it is also obvious that organisms bring about changes in environments. Numerous animals manufacture nests, burrows, holes, webs and pupal cases. Plants change the levels of atmospheric gases and modify nutrient cycles. Fungi decompose organic matter, and bacteria engage in decomposition and nutrient fixation. The standard view of evolution does not deny this, but treats niche construction as no more than the product of selection.

Conversely, from the niche-construction perspective, evolution is based on networks of causation and feedback. Organisms drive environmental change and organism-modified environments subsequently select organisms. The argument that niche construction does not play a causal role in evolution because it is partly a product of natural selection, makes no more sense than would the counter-proposal that natural selection can be disregarded because it is partly a product of niche construction.

At present, standard evolutionary theory models the evolutionary consequences of niche construction solely in terms of fitness ‘pay-offs’ to the genes expressed in niche construction. For us, this is unsatisfactory. First, it



Networking: do birds' nests shape evolution?

misses part of the causal story. When a beaver builds a dam it not only affects the propagation of dam-building genes, but it must also transform the selection acting on a host of other beaver traits. Second, some organism-driven changes in the environment persist as a legacy to modify selection on subsequent generations — an ecological inheritance. Contemporary earthworms are adapting to a soil environment largely constructed by their ancestors. Third, genes often exert a weak regulatory influence on niche construction. There are no ‘genes for’ dairy farming; nor is it (strictly) an adaptation, but rather an adaptive cultural practice. Yet this activity has generated selection that favours genes for lactose absorption.

The niche-construction perspective was brought to the attention of evolutionary biologists by Richard Lewontin in the 1980s. The classic view of evolution, he said, can be encapsulated in the pair of differential equations $O' = f(O, E)$ and $E' = g(E)$, where the primes indicate differentiation with respect to time, the state variables are the organism (O) and environment (E), and the function f stipulates evolution (g , the function describing environmental change, is considered beyond the remit of evolutionary biology). Lewontin claimed a more accurate formulation to be $O' = f(O, E)$ and $E' = g(O, E)$, because part of the cause of evolution is the impact of the organism on its environment. Yet his seminal essays had little immediate impact. What has changed?

First, accumulated empirical evidence for niche construction, in everything from ants to boreal forests, means that there is no longer any dispute over its ubiquity. The view of leading evolutionist Theodosius Dobzhansky that “man alone adapts himself ... by actively or even deliberately changing the environment”

is no longer tenable. Niche construction is not the prerogative of large populations, keystone species or clever animals — it is a fact of life.

Second, recent mathematical theory has established that niche construction changes the evolutionary dynamic. Niche construction can create new equilibria, affect the stability of others, generate unusual phenomena, such as momentum effects (where populations continue to evolve in the same direction after selection has stopped or reversed) and inertia effects (a delayed evolutionary response to selection), as well as opposite and catastrophic responses to selection.

Third, and most importantly, empirical methods have been devised that allow evolutionary biologists and ecologists to ask — and answer — new questions and shed fresh light on old ones. For instance, evolutionary biologists can use genomics or proteomics techniques to track the evolutionary consequences of niche construction in microorganisms such as bacteria. And ecologists can make predictions about which species will invade a community and whether or not the invasion will be benign, based on knowledge of the invader's tolerance to the residents' niche construction, and the nature of the invader's activities.

Already, niche construction is stimulating palaeobiologists and ecologists to organize conferences and symposia, philosophers to commission special editions of journals and researchers as diverse as archaeologists and primatologists to grow excited by its ideas and methods. Ultimately, it is here that the fate of the niche-construction ‘school’ will be decided. If this early enthusiasm proves a fleeting fad, and there is little genuine utility in the alternative panorama, it will fade. However, if researchers continue to start using these new tools then, in the words of philosopher David Hull: “the result should be a massive reorientation of evolutionary theory”.

Kevin N. Laland is in the School of Biology, St Andrews University, UK. John Odling-Smee is at the Institute of Biological Anthropology, Oxford University, UK. Marcus W. Feldman is in the Department of Biological Sciences, Stanford University, USA.

FURTHER READING

- Jones C. G. *et al.* *Ecology* **78**, 1946–1957 (1997).
 Laland K. N., Odling-Smee J. & Feldman M. W. *Proc. Natl Acad. Sci. USA* **96**, 10242–10247 (1999).
 Odling-Smee J., Laland K. N. & Feldman M. W. *Niche Construction. The Neglected Process in Evolution. Monographs in Population Biology* **37** (Princeton Univ. Press, Princeton, 2003).
 Vandermeer J. *Science* **303**, 472–474 (2004).
 ▶ www.nicheconstruction.com

A great grand-daddy of ice cores

Jerry F. McManus

A record of Earth's climate over the past eight ice ages and their associated interglacial periods has been uncovered from a new ice core in Antarctica, almost doubling the age of previous ice-core records.

Millennium after millennium, the snow falling on Greenland and Antarctica has built up deep ice sheets that are invaluable archives of past conditions on Earth. Antarctica has now yielded the longest ice-core record yet¹, one that covers a staggering 740,000 years, with more to come. This accomplishment is the result of an effort by the EPICA consortium (European Project for Ice Coring in Antarctica). The group's report, by lead author Eric Wolff and colleagues, appears on page 623, with further comment on page 596.

A decade ago, ice cores from Greenland provided compelling evidence for the persistent climatic instability of the last glacial cycle², which reached back more than 100,000 years through the last ice age and into the previous interglacial. In the polar desert of Antarctica, however, slower snow accumulation means the ice archives can span longer intervals, and what is arguably the single best climate record in existence comes from Vostok, East Antarctica³. This goes back approximately 400,000 years, taking in four glacial cycles. Yet even this grand-daddy of ice cores falls frustratingly short of reaching several climatic milestones — for instance, the interval when the dominant behaviour of the ice ages shifted; the time of the largest deglacial transition; and the early portion of an interglacial interval that began shortly before 400,000 years ago and may be the best analogy to the current interglacial that began more than 10,000 years ago⁴.

Wolff and colleagues¹ report the first results obtained by drilling a core from 3-km-thick ice at a site known as Dome C in East Antarctica. These results provide a spectacular record that extends back through eight glacial cycles (Fig. 1), confirming aspects of the Vostok record and providing a view of the preceding climatic milestones.

A large team is necessary in projects of this sort because of the demanding field operations, and also because of the many analytical approaches needed to extract information from the ice. For example, measurements of the physical characteristics of the atmosphere over the ice sheet (such as its temperature and dustiness recorded in snow and eventually ice) may be directly compared with data about the composition of the atmosphere itself (recorded in the air bubbles trapped as snow turns to ice). Such comparisons have established the connection

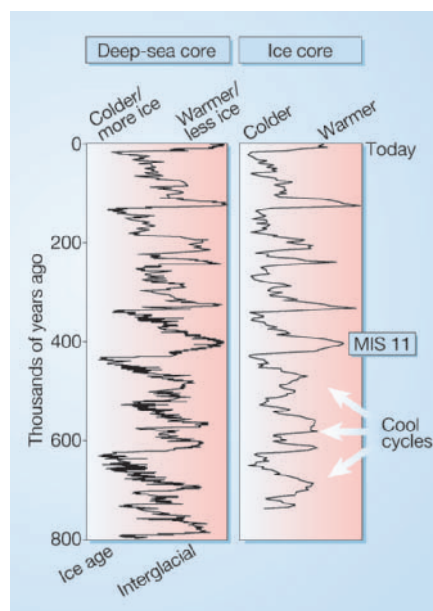


Figure 1 Glacial cycles of the past 800,000 years. The history of deep-ocean temperatures and global ice volume is inferred from a high-resolution record of oxygen-isotope ratios measured in bottom-dwelling foraminifera shells preserved as microfossils in Atlantic Ocean sediments^{6,10,11}. Air temperatures over Antarctica are inferred from the ratio of deuterium and hydrogen in the ice at Dome C (ref. 1). Marine Isotope Stage 11 (MIS 11) is an interglacial that bears some similarity to the most recent warm period⁴, yet lasted many thousands of years longer. Note the smaller magnitude and less-pronounced interglacial warmth of the glacial cycles that preceded MIS 11.

between changes in atmospheric concentrations of carbon dioxide and temperature during the past few glacial cycles, and a similar connection between increased dust deposition and glacial periods that are thus inferred to be arid and/or windy.

The Dome C record includes dust concentrations and temperature estimates based on the deuterium to hydrogen ratio in the ice. Further work will increase the resolution of these data, and extend them to the base of the ice, perhaps reaching back another 200,000 years. Greenhouse-gas data have been generated for the crucial interval around 400,000 years ago, and these will also be extended back in time.

Ice cores are not the only source of climate records. They are complemented by data

from terrestrial settings and from deep-sea sediments, which are especially useful because they provide better spatial coverage of the Earth and may extend much farther back in time. Both deep-sea and ice cores provide evidence for local temperature changes, ice cores additionally providing indications of atmospheric composition through the trapped air bubbles. Sediment cores contain the global signature of each glaciation, as recorded in the oxygen-isotope ratio of buried microfossil shells. This signature in part reflects temperature changes and in part the amounts of water residing in the isotopically very different reservoirs for water on Earth's surface — ice and ocean. It constitutes the basis for the scheme of marine isotope stages (MIS) for numbering glacial cycles, which counts the alternating ice ages and interglacials back in time, beginning with the present interglacial, MIS 1.

An intriguing aspect of long-term climate, revealed by deep-sea cores, is a shift in the pattern of glacial cycles that occurred without changes in the varying seasonal distribution of sunlight that is believed to pace the ice ages. Before one million years ago, glaciations of moderate extent waxed and waned in concert with 40,000-year variations in the tilt of Earth's rotation axis. During the past half-million years, the dominant period of these cycles is more like 100,000 years, similar to variations in the shape of Earth's elliptical orbit. So a transition in Earth's climatic response must have occurred as it entered this '100K world'. These longer climate cycles have also been the largest in magnitude of all the ice-age cycles, combining with a long-term climatic trend towards more glacial conditions over millions of years to produce several of the most extreme glaciations on record (the last glacial maximum of the last ice age, MIS 2, for instance). Remarkably, the 100K cycles have also resulted in the four mildest interglacial intervals of the past million years — including MIS 1, which has, of course, seen the advent of human civilizations.

The new record¹ provides further information about this transition, with the promise of more to come. Before 450,000 years ago, the glacial cycles display smaller amplitudes. Interestingly, the glacial portion of each cycle evinces nearly the same amount of cooling as during the more recent ice ages, whereas the warmer portion of each earlier

cycle is clearly less pronounced. Wolff and colleagues also note that although the older cycles do not achieve nearly the same degree of warmth as subsequently, the warm intervals represent a much greater portion of the total cycle than was the case for the past few peak interglacial times; these seem to be increasingly brief, albeit pronounced, respites from the more typical harsh, unstable climate of the 100K world⁵.

The EPICA team's recovery of undisturbed ice representing all of MIS 11 is of especial note. When this interglacial occurred, some 400,000 years ago, the slowly varying shape of Earth's elliptical orbit around the Sun was nearly circular. In this configuration, variations in the seasonal distribution of sunlight are due primarily to the tilt of Earth's axis, with little influence of the precession of the seasons around the orbit every 20,000 years. Because the Sun occupies a position that is nearer to one end of the ellipse, this precession results in the gradual cooling of northern summers over 10,000 years, as they migrate towards the more distant end of the orbit, culminating in conditions that are most conducive to year-round snow cover and the initiation of an ice age. A similar orbital configuration to that for MIS 11 holds today, and will for some time, making this interglacial a potential analogue for the natural development of Earth's climate in the future⁴. Northern Hemisphere summer now occurs when the Earth is near its farthest point from the Sun, although the nearly circular orbit raises the question of whether the resulting weak decline in northern summer insolation of today was sufficient to trigger the onset of an ice age in the past⁶.

Wolff *et al.* confirm that concentrations of the greenhouse gases CO₂ and CH₄ were also similar to pre-industrial levels during MIS 11, showing that the overall controls on incoming and outgoing radiation that drive climate were similar. Their deuterium and dust records indicate that the magnitude of the climatic response was likewise similar, and the 28,000 years of relative warmth in Antarctica is in keeping with evidence for prolonged warmth elsewhere^{6–10}. This suggests that Earth's climate may have 'skipped a beat' when its orbit was nearly circular⁶, rather than returning to an ice age after the 10,000 years of one half-precession cycle. The new results are perhaps the best evidence yet that MIS 11 was akin to the current interglacial — except, of course, that we now have rising concentrations of CO₂ and have yet to see if 'our' interglacial might last another 20,000 years.

A priority for further investigations will be completing the gas records from air trapped in the Dome C ice. Given the complete record, it may be possible to detect an overall decline in greenhouse gases that might help to account for the long-term

deterioration of global climate. Equally important is the question of whether the earlier warm intervals were associated with lower greenhouse-gas concentrations than have been achieved within the past few cycles. Such evidence will help us to find out whether the accentuated warmth of peak interglacial intervals within the 100K world was associated with 'elevated' atmospheric greenhouse-gas levels that have been increasingly surpassed in the post-industrial era. ■

Jerry F. McManus is in the Department of Geology and Geophysics, 121 Clark Laboratory, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA.

e-mail: jmcmanus@whoi.edu

Animal behaviour

Eavesdropping on bats

Brock Fenton and John Ratcliffe

Two investigations into bat echolocation provide striking examples of the sophistication and the possible evolutionary and ecological consequences of variability in call design.

In 1794, Lazzaro Spallanzani reported experimental results supporting his earlier proposal that bats could 'see' with their ears. The famed Georges Cuvier found the suggestion preposterous¹, however, and it took almost another 150 years for Spallanzani to be vindicated. After repeating many of Spallanzani's experiments, Donald Griffin² published the same conclusions in 1940, coining the term 'echolocation' to describe how bats use echoes of the sounds they produce to locate objects in their path. A microphone sensitive to sound frequencies above the range of human hearing, a bat detector, allowed Griffin to eavesdrop on what bats said as they flew through an obstacle course in the dark.

Today, we know that there is variation between bat species in the design of echolocation calls, which often coincides with differences in their behaviour and ecology³. Two papers in this issue, by Kingston and Rossiter⁴ and Siemers and Schnitzler⁵, advance this line of investigation further.

Kingston and Rossiter (page 654)⁴ examined the situation in a single species, the large-eared horseshoe bat (*Rhinolophus philippinensis*), which occurs from southeast Asia to Australia. They showed how echolocation signals can diverge within a species and how this divergence might promote sympatric speciation — the division of one species into two or more without a geographical barrier. This is a hot and contentious topic in evolutionary biology. In three study areas, Kingston and Rossiter found three distinct variants of large-eared horseshoe bats differing in size, echolocation calls and relatedness. The largest was almost

1. EPICA community members *Nature* **429**, 623–628 (2004).
2. Dansgaard, W. *et al.* *Nature* **364**, 218–220 (1993).
3. Petit, J. R. *et al.* *Nature* **399**, 429–436 (1999).
4. Loutre, M. F. & Berger, A. *Glob. Planet. Change* **36**, 209–217 (2003).
5. McManus, J. F., Oppo, D. W. & Cullen, J. L. *Science* **283**, 971–975 (1999).
6. McManus, J. F., Oppo, D. W., Cullen, J. L. & Healey, S. L. in *Earth's Climate and Orbital Eccentricity: The Marine Isotope Stage 11 Question* (eds Droxler, A., Poore, R., Burckle, L. & Osterman, L.) 69–85 (Am. Geophys. Un., Washington DC, 2003).
7. Winograd, I. J., Landwehr, J. M., Ludwig, K. R., Coplen, T. B. & Riggs, A. C. *Quat. Res.* **48**, 141–154 (1997).
8. Poore, R. Z. & Dowsett, H. J. *Geology* **29**, 71–74 (2001).
9. Hodell, D. A., Charles, C. D. & Ninnemann, U. S. *Glob. Planet. Change* **24**, 7–26 (2000).
10. Oppo, D. W., McManus, J. F. & Cullen, J. L. *Science* **279**, 1335–1338 (1998).
11. Flower, B. P. *et al.* *Paleoceanography* **13**, 388–403 (2000).

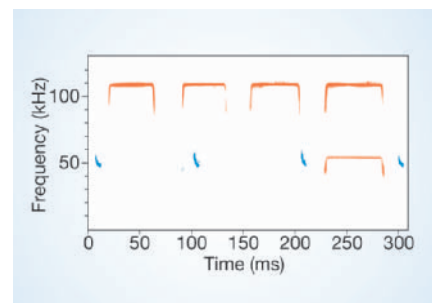


Figure 1 The two sides of the echolocation fence. These calls represent bats that separate pulse and echo in frequency (*Rhinolophus*, high-duty cycle, orange) and those separating them in time (*Pipistrellus*, low-duty cycle, blue). The high-duty calls are long, separated by short periods of silence and dominated by a single frequency; the low-duty calls are short, separated by long periods of silence and are not dominated by a single frequency. The fourth call of the *Rhinolophus* includes a lower-frequency harmonic as discussed by Kingston and Rossiter⁴. The bats foraged along the edge of a watercourse south of the Dead Sea, and the recording — made by M. B. Fenton — is unusual because it captures both types of cycle simultaneously.

twice as heavy as the smallest, and the sounds dominating their echolocation calls ranged from 27.2 ± 0.2 kHz in the largest to 53.6 ± 0.6 kHz in the smallest.

The level of detail available to an echolocating bat is a function of the wavelength of the sounds in its echolocation calls, and so differences in the frequencies that dominate its calls influence a bat's auditory scene⁶. Bats using high frequencies (shorter wavelengths) can detect smaller prey than can

bats using lower-frequency calls (longer wavelengths). Kingston and Rossiter suggest that the range of echolocation calls in one species would generate 'disruptive selection' because larger bats do not have the same access to small prey as do smaller ones. Theirs is the first demonstration of how adaptive evolution in bats, and so speciation, might have been driven through divergences in echolocation signals.

For their part, Siemers and Schnitzler (page 657)⁵ examined the behavioural consequences of differences in echolocation signals used by similar species of bats to detect prey. In a portable flight-room, they challenged flying individuals of five European species of mouse-eared bats (*Myotis* species) to detect and attack prey sitting on or close to vegetation. This is presumed to be difficult for the bats because echoes from prey could be masked by echoes — 'clutter' — from the background. Siemers and Schnitzler standardized the degree of clutter in which the bats operated, and documented their behaviour and foraging performance. The five species they used have similar hunting behaviour and are placed in the same 'foraging guild' of bats (the 'edge space aerial/trawling foragers'). The five species might have been expected to perform at the same level, but they did not.

In the tradition of Griffin and Spallanzani, Siemers and Schnitzler controlled for other cues (vision, olfaction) and demonstrated a significant relationship between the design of echolocation calls and foraging performance. Specifically, they showed that foraging performance in clutter was predictable from echolocation call design, particularly from differences in calls that had been considered minor. Their study is the first to provide empirical evidence that seemingly minor differences in call design can have real behavioural consequences. In contrast to Kingston and Rossiter, Siemers and Schnitzler show that signal designs of similar species can converge, reflecting foraging behaviour that is independent of presumed evolutionary relationships.

Individually and jointly, these two papers advance our understanding of the diversity of echolocation in bats. They have opened doors to a better appreciation of the variety of echolocation call designs, including the identification of cryptic species⁷ — that is, the discovery that what had been considered a single species really consists of two or more. Coupled with data on the enhanced echoes that some flowers return to the bats that pollinate them⁸, the new findings also allow better interpretation of insights into other pressures acting on the evolution of bats. For example, another component of the echolocation story is the listeners — other bats, or other animals that, like Griffin, eavesdrop on the calls^{9,10}. Kingston and Rossiter's work shows clearly that changes in echolocation calls can affect

not only bats' views of the world, but also the ability of one individual to communicate with another. Siemers and Schnitzler's results set the stage for examining the influence of call design on the ability of potential insect prey to detect and evade hunting bats^{9,10}.

The two studies^{4,5} used species from both sides of the bat echolocation fence. On one side, mouse-eared bats, like most bats, separate call and echo in time (low-duty cycle); on the other, horseshoe bats separate them in frequency (high-duty cycle) (Fig. 1). Both approaches to echolocation are ancient, with fossil evidence indicating that they were present in bats some 50 million years ago¹¹. The new data speak to the divergence of call design after the evolution of echolocation, but the early history of bats and echolocation remains unclear. There is plenty of opportunity in this line of research: stay tuned for the next chapter. ■

Particle physics

From the top...

Georg Weiglein

The top quark is by far the heaviest elementary particle known. A measurement of its mass with higher precision has bearing on our understanding of the fundamental interactions of nature.

The basic building-blocks of matter, as far as we know, are quarks and leptons, together with the force-carrying particles that mediate their interactions. Quarks and leptons (the latter group including the electron) are grouped in three generations; the particles in the second and third generations seem a perfect copy of those of the first generation, except that their masses are much larger. The top quark is the heaviest of all quarks and leptons, and is central to some of the most pressing questions in particle physics. For instance, why is the third-generation top quark more than 300,000 times heavier than the first-generation electron? Why are there two other quarks with precisely the same properties as the top quark but with very different masses? And what is the origin of mass itself?

Precise knowledge of the mass of the top quark and its interactions is a key ingredient in testing theory against experimental data. On page 638 of this issue¹, the DØ Collaboration report an improved measurement of the top-quark mass, using data taken at the Tevatron proton-antiproton collider at Fermilab, near Chicago. Combining this with previous measurements from DØ and its sister experiment CDF, the new world average² for the mass of the top quark is 178.0 ± 4.3 GeV/ c^2 , where c is the speed of light (the mass of the proton expressed in these units is about 1 GeV/ c^2). Compared with the previous world average³, the central value of the mass has shifted

Brock Fenton is in the Department of Biology, University of Western Ontario, London, Ontario N6A 5B7, Canada.

e-mail: bfenton@uwo.ca

John Ratcliffe is in the Department of Zoology, University of Toronto at Mississauga, Ontario L5L 1C6, Canada.

e-mail: j.ratcliffe@utoronto.ca

1. Griffin, D. R. *Listening in the Dark* (Yale Univ. Press, New Haven, 1959).
2. Galambos, R. & Griffin, D. R. *Anat. Rec.* **78**, 95 (1940).
3. Thomas, J., Moss, C. & Vater, M. (eds) *Echolocation in Bats and Dolphins* (Univ. Chicago Press, 2004).
4. Kingston, T. & Rossiter, S. J. *Nature* **429**, 654–657 (2004).
5. Siemers, B. M. & Schnitzler, H.-U. *Nature* **429**, 657–661 (2004).
6. Simmons, J. A. & Stein, R. A. *J. Comp. Physiol. A* **135**, 61–84 (1980).
7. Barratt, E. M. et al. *Nature* **387**, 138–139 (1997).
8. von Helversen, D. & von Helversen, O. *Nature* **398**, 759–760 (1999).
9. Pye, J. D. *Nature* **218**, 797 (1968).
10. Fullard, J. H. in *Comparative Hearing: Insects* (eds Hoy, R. R., Popper, A. N. & Fay, R. R.) 279–326 (Springer, New York, 1998).
11. Simmons, N. B. & Geilser, J. H. *Bull. Am. Mus. Nat. Hist.* **235**, 1–182 (1998).

upwards by about 4 GeV/ c^2 . The experimental error has been reduced by about 15%, sharpening our view of the underlying physics.

The role of the top quark in disentangling the fundamental principles of nature is twofold. On the one hand, its large mass makes the top quark a prime target in the search for new physics that might so far be unaccounted for. For instance, the long-hypothesized Higgs boson, which is the last missing ingredient of the standard model of particle physics, is predicted to interact with other particles with a strength that is proportional to their masses. So the physics of the heavy top quark would be significantly influenced by its interaction with the Higgs boson. On the other hand, the mass of the top quark is a key parameter in the predictions for many observable quantities. Small deviations between measurement and prediction could be a signal of new physics, so the uncertainty in the predictions that arises from the experimental error on the top-quark mass limits the sensitivity of experiment to new physics.

The values of several precisely measured quantities, as predicted by the standard model, depend on the square of the top-quark mass, M_t ; their dependence is much weaker on the as yet unknown mass of the Higgs boson (so far, experiment has excluded⁴ any mass value below 114.4 GeV/ c^2). Therefore, in using a so-called global fit of the model predictions to all available data, an improved knowledge of M_t better constrains



100 YEARS AGO

An interesting mathematical study of the conditions which probably obtained in the primitive solar nebula has been communicated to the Academy of Science of St. Louis by Mr. Francis E. Nipher... According to the equations developed by the author, it seems impossible that at the time when the planets were separating from the parent mass the nebula was wholly gaseous. The idea that the planets were formed from condensing swarms of meteorites is the only reasonable one which conforms with the numerical results obtained. It also appears that at the times when the moon separated from the earth, and Mercury from the sun, the respective parent masses must have been in the solid state, the sun having fused and become vaporised since the separation of Mercury. Further, it seems unnecessary, and even improbable, that the earth should ever have been in a state of fusion. By substituting the proper conditions in one of his general equations, Mr. Nipher finds that the isothermal 7000° C. is probably the one existing at the sun's surface at the present time.
From *Nature* 9 June 1904.

50 YEARS AGO

A variety of cells and tissues of mammals survive for long periods at the temperature of 'dry ice', - 79° C., when frozen in media containing glycerol. At the other extreme, Andjus's work on the whole animal shows that, by special methods of cooling and rewarming, rats can be revived from deep body temperatures of about + 0.5° C.... Hamsters, chosen because of their known adaptability to body temperatures between 2.5 and 38° C., were cooled by the method recently described by Andjus and Smith for rats... The animals stiffened progressively during this process until they were wood-like to the touch, and presumably consisted of a hard frozen shell surrounding an unfrozen or only slightly frozen interior... The extremities intimately exposed to the bath fluid at - 4° C. to - 7° C. were severely frozen — the ears, for example, attained the consistency of cardboard, and may well have undergone crystallization of 80 per cent of their water content. It is remarkable, however, that damage obviously due to this cause has been seen in only two of the twenty-one animals revived completely, some of which have been kept for many weeks afterwards in apparently normal health.
A. U. Smith, J. E. Lovelock, A. S. Parkes
From *Nature* 12 June 1954.

the likely value of the Higgs-boson mass. In fact, the 4 GeV/c² shift in the central value of M_t has shifted the upper limit on the Higgs-boson mass by more than 30 GeV/c², to 251 GeV/c² (at 95% confidence level)¹. This upper limit has an important impact on the experimental strategies used to search for the Higgs boson at present and future colliders.

Finding the Higgs boson in the predicted range would be another triumph for the standard model — as, of course, was the discovery of the top quark itself. Historically, the mass of the top quark had been predicted from a global fit to a wealth of precise measurements made at the LEP and SLC electron-positron colliders (at CERN in Geneva and at SLAC in Stanford, respectively). The top quark was discovered at the Tevatron in 1995, with a mass value in perfect agreement with the predicted range.

Although the standard model has passed many experimental tests with great success, it cannot be the ultimate theory of the fundamental interactions. This is evident from the fact that it describes only three of the four known interactions — namely, the electromagnetic, weak and strong interactions, but not gravity. It also has several theoretical shortcomings and leaves many questions unanswered. Perhaps the most attractive framework for extending the standard model is supersymmetry. A supersymmetric extension of the standard model could be the low-energy limit of a more fundamental high-energy theory that would consistently include gravity and would describe all the fundamental forces in a unified way. Supersymmetric theories predict that there are partners for all the known particles. The minimal supersymmetric extension of the standard model — the 'MSSM' — comprises one pair of superpartners for each quark and lepton, superpartners for the force carriers, and five Higgs bosons.

In supersymmetric models, as a consequence of the higher degree of symmetry, the mass of the lightest Higgs boson can be predicted directly (in contrast to the standard model, in which the Higgs mass is a free parameter, allowing only an indirect determination via a global fit). The predicted mass is very sensitive to the mass of the top quark, scaling as M_t^4 — an even more pronounced dependence than in the standard-model case. Figure 1 shows the prediction^{5,6} for the lightest Higgs-boson mass in the MSSM: the effect of the change in the top-quark mass, to 178.0 ± 4.3 GeV/c², is clearly seen. The direct experimental detection of the Higgs boson would enable its mass to be measured with an accuracy below the 1% level. Thus, a precise knowledge of M_t with an accuracy even better than presently available will be crucial for Higgs physics in supersymmetric extensions of the standard model⁷.

Besides having an important impact on Higgs physics, the top-quark mass influences

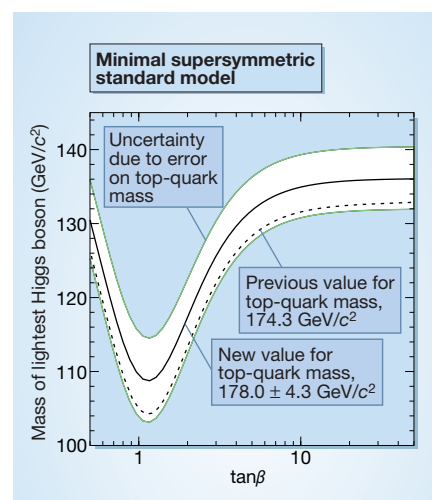


Figure 1 The mass of the lightest Higgs boson in the minimal supersymmetric standard model (MSSM). The predicted value^{5,6} is shown as a function of the parameter $\tan\beta$, which relates the properties of the different Higgs bosons of the MSSM to each other (the other MSSM parameters are chosen such that they maximize the resulting value of the Higgs mass). The predicted Higgs mass is sensitive to the value of the top-quark mass used in the calculation. The solid line indicates the prediction using the new measurement of the top-quark mass from the DØ Collaboration¹; the white band indicates the uncertainty of the prediction that results from the error on the top-quark mass. The dashed line shows the situation before the new measurement (the previous experimental error of ± 5.1 GeV/c² is not shown). Based on the new value of the top-quark mass, an upper bound on the mass of the lightest MSSM Higgs boson of about 140 GeV/c² is established.

many other predictions of the MSSM — for instance, the masses of the superpartners of the top quark and the strengths of their interactions. The ultimate goal is to connect the predictions of the MSSM, or other extensions of the standard model, with a more fundamental theory at a higher energy scale. This may provide evidence for the unification of all of the forces of nature into a single fundamental interaction. Measurements made at the energy scales directly accessible to us in collider experiments can be extrapolated to very high energy scales, but for this to be reliable a precise knowledge of M_t is crucial⁷. If the extrapolation is sufficiently precise, it may even give us clues about the structure of the unified force itself.

Further progress will require new experimental data — both the discovery of new particles, such as the Higgs boson or supersymmetric partners, and more precise measurements of observable quantities that allow a sensitive test of the underlying theory. Among these, improving the accuracy of the measurement of the top-quark mass will continue to be of the utmost importance. From data taken during the present phase

of operation at the Tevatron (known as 'Run II'), the experimental error on M_t will be reduced to 2–3 GeV/ c^2 ; at the Large Hadron Collider⁸, currently under construction at CERN, this accuracy will be improved further to 1–2 GeV/ c^2 .

The ultimate precision on M_t , however, will be achieved at a linear electron–positron collider. Such a machine is currently in the planning phase and could go into operation around the middle of the next decade. Data from the linear collider could improve the accuracy on the top-quark mass by about a factor of ten^{9–11}. Only then will the uncertainty due to the experimental error of the top-quark mass be well enough under control for the information gleaned from the LHC in the next decade — on the Higgs boson (or bosons), supersymmetric partners or other new physics — to be fully exploited. ■

Georg Weiglein is at the Institute for Particle Physics

Phenomenology, Department of Physics,
University of Durham, Durham DH1 3LE, UK.
e-mail: georg.weiglein@durham.ac.uk

1. The DØ Collaboration *Nature* **429**, 638–642 (2004).
2. The CDF Collaboration, the DØ Collaboration and the Tevatron Electroweak Working Group. Preprint at <http://arxiv.org/abs/hep-ex/0404010> (2004).
3. Hagiwara, K. *et al.* *Phys. Rev. D* **66**, 010001, 271–433 (2002).
4. The ALEPH, DELPHI, L3 and OPAL Collaborations and the LEP Working Group for Higgs Boson Searches *Phys. Lett. B* **565**, 61–75 (2003).
5. Heinemeyer, S., Hollik, W. & Weiglein, G. *Comput. Phys. Commun.* **124**, 76–89 (2000).
6. www.feynhiggs.de
7. Heinemeyer, S., Kraml, S., Porod, W. & Weiglein, G. *J. High Energy Phys.* JHEP09(2003)075 (2003).
8. Beneke, M. *et al.* in *Standard Model Physics (and More) at the LHC* (eds Altarelli, G. & Mangano, M.) 419–529 (CERN, Geneva, 1999).
9. Heuer, R.-D. *et al.* Preprint at <http://arxiv.org/abs/hep-ph/0106315> (2001).
10. Abe, T. *et al.* Preprint at <http://arxiv.org/abs/hep-ex/0106056> (2001).
11. ACFA Linear Collider Working Group. Preprint at <http://arxiv.org/abs/hep-ph/0109166> (2001).

Interstellar chemistry

Molecular nitrogen in space

Theodore P. Snow

Astronomers have found evidence of molecular nitrogen in the clouds of gas between the Earth and a distant star. The chemistry involved in the formation of these diffuse clouds might need to be rethought.

On page 636 of this issue, David C. Knauth and colleagues¹ claim the first detection of molecular nitrogen (N_2) in interstellar space. This simple diatomic molecule, made of one of the most abundant elements in the Universe, is the most common constituent of Earth's modern atmosphere. It is also a major component of the atmosphere of Saturn's moon Titan, and has been detected in trace amounts in the atmospheres of Venus and Mars. But it has proved surprisingly difficult to find N_2 in any environment beyond the Solar System.

Chemical models of dark interstellar clouds (whose densities are usually in the range of 10^3 to 10^5 particles per cm^3) suggest that N_2 should be the most abundant form of nitrogen in these regions. This leads to the prediction^{2–4} that the ratio of N_2 to hydrogen should be about 10^{-5} . In contrast, models for diffuse interstellar clouds, which are transparent and have densities of about 10^2 particles per cm^3 , predict a much lower N_2 abundance, in the range between 10^{-9} and 10^{-8} that of hydrogen^{2,5}.

Both predictions suggest that N_2 might be observable, but searches for this molecule in interstellar space had been fruitless until now. One of the difficulties in detecting interstellar N_2 arises from the fact that the symmetric diatomic molecule has no allowed rotational or vibrational (dipole) transitions. Thus, N_2 — unlike most of the

120 or more species now detected in dark interstellar clouds — cannot be detected either through millimetre-wavelength observations of rotational emission lines or through infrared spectroscopic detection of vibrational bands (absorption or emission).

The only viable approach to finding interstellar N_2 is to search for the spectral lines created by electronic transitions in the molecule. These lines are found exclusively at far-ultraviolet wavelengths (shorter than 100 nm), for which space-based telescopes are required because the Earth's atmosphere blocks such radiation. For technical reasons, however, most ultraviolet telescopes have not covered the far-ultraviolet spectral region where the N_2 bands lie. For example, the Hubble Space Telescope cuts off at about 115 nm, well above the wavelength needed for an N_2 search. The Copernicus satellite — a small mission that was developed and led by the late Lyman Spitzer and operated from 1972 until 1980 — was the first orbiting spectroscopic observatory capable of far-ultraviolet searches for N_2 in interstellar space, but no detection was achieved⁶.

The best chance for astronomers to search for interstellar N_2 has been afforded by the Far Ultraviolet Spectroscopic Explorer (FUSE) mission, now in its fifth year of operation. FUSE was designed specifically to extend ultraviolet spectroscopy to the shorter wavelengths that are not accessible to the Hubble Space Telescope, including

the spectral region where the electronic bands of N_2 lie. Knauth *et al.*¹ have taken advantage of FUSE's far-ultraviolet sensitivity to search for N_2 — and apparently they have found it.

In a classic example of spectroscopic sleuth work, Knauth *et al.* have detected absorption by N_2 in the line of sight towards the star HD 124314 by sorting through and eliminating other features that are blended into the spectrum. These other features arise through the absorption of radiation by the star's own atmosphere, by foreground interstellar gas (mostly molecular hydrogen) and by N_2 in the outer vestiges of Earth's atmosphere. The detection of interstellar N_2 was aided by the fact that several individual N_2 lines are accessible to FUSE and also because FUSE covers the N_2 wavelength region with two separate detectors, which means that instrumental artefacts in the data can be eliminated.

The line of sight towards HD 124314 does not intersect a dark molecular cloud; rather, this is a long pathlength, probing one or more diffuse clouds. So, according to model calculations, the ratio of N_2 to hydrogen should be closer to the 10^{-9} to 10^{-8} level that is predicted for diffuse clouds than to the 10^{-5} level predicted for dense clouds. Knauth *et al.* have found an intermediate value, with N_2 representing about 10^{-7} of the total hydrogen abundance in their observed line of sight. This abundance of N_2 does not fit either the dense-cloud or diffuse-cloud models.

Among the possible explanations is that the line of sight towards this star contains one or more 'translucent' clouds, which are reckoned by astronomers to be intermediate (or possibly transitional) between dense and diffuse clouds⁷. Alternatively, the models for diffuse clouds might be incorrect, or the detection claimed by Knauth *et al.* is wrong. The first and third of these options can probably be eliminated, as the line-of-sight dust extinction to this particular star is too small to include a translucent cloud, and the claimed detection of N_2 seems secure. So we must surmise that the chemical models for N_2 in diffuse clouds are inadequate.

Normally it is assumed that, with the exception of hydrogen, molecules in diffuse clouds form through gas-phase chemical reactions^{2,5}. But in dense clouds an additional process, that of molecule formation on grain surfaces, is probably important^{8,9}. The detection of N_2 by Knauth *et al.*¹ suggests that grain-surface reactions might contribute more to diffuse-cloud chemistry than previously thought. This conclusion is consistent with earlier searches in diffuse clouds for NH, another simple diatomic molecule found to be more abundant than expected from gas-phase chemistry alone^{10,11}. If grain-surface reactions are required to explain the measured abundances of N_2 and NH, it is possible that other surface reactions

should also be incorporated into models of the chemistry inside these diffuse clouds. ■

Theodore P. Snow is at the Center for Astrophysics and Space Astronomy, University of Colorado, Boulder, Colorado 80309-0389, USA.
e-mail: tsnow@casa.colorado.edu

1. Knauth, D. C., Andersson, B.-G., McCandless, S. R. & Moos, H. W. *Nature* **429**, 636–638 (2004).
2. Viala, Y. P. *Astron. Astrophys. Suppl.* **64**, 391–437 (1986).
3. Womack, M., Ziurys, L. M. & Wyckoff, S. *Astrophys. J.* **393**, 188–192 (1992).

4. Bergin, E. A., Langer, W. D. & Goldsmith, P. F. *Astrophys. J.* **441**, 222–243 (1995).
5. Black, J. H. & Dalgarno, A. *Astrophys. J. Suppl.* **34**, 405–423 (1977).
6. Lutz, B. L., Owen, T. & Snow, T. P. *Astrophys. J.* **227**, 159–162 (1979).
7. van Dishoeck, E. F. & Black, J. H. *Astrophys. J.* **340**, 273–297 (1989).
8. Millar, T. J. *Mon. Not. R. Astron. Soc.* **199**, 309–319 (1982).
9. Ehrenfreund, P. & Charnley, S. B. *Annu. Rev. Astron. Astrophys.* **38**, 427–483 (2000).
10. Crutcher, R. M. & Watson, W. D. *Astrophys. J.* **209**, 778–781 (1976).
11. Meyer, D. M. & Roth, K. C. *Astrophys. J. Lett.* **376**, L49–L52 (1991).

Economics

The wealth of nations

Jared Diamond

A country's affluence depends partly on its institutions. Geographic and other factors yield a fuller explanation, illuminate the origins of 'good' institutions, and suggest targets for foreign aid.

Why are per capita income and gross national product more than 100 times higher in some countries than in others? Why are resource-poor Iceland and Luxembourg among the ten richest countries, while resource-rich Bolivia and Nigeria are among the poorest? This question is of practical as well as academic interest: if we knew the answers, perhaps poor countries could use them to achieve wealth. But the question is controversial and complicated, as a burst of publications attests^{1–9}.

The usual view is that differences in national wealth arise from differences in the accumulation of physical and social capital and in the adoption of new technology, due in turn to differences in the quality of political and economic institutions^{10–14}. Rich countries are rich because they have 'good' institutions promoting investment and accumulation of wealth. A conclusion from this view, embraced by many aid programmes, is that the best way to help poor countries is to assist them in developing good institutions — such as the rule of law, honest, efficient government, impartial enforcement of contracts, unimpeded flow of capital and goods across international borders, and protection of investors' property. The most convincing support for this view comes from four pairs of neighbouring countries sharing the same environment, one of them rich and with 'good' institutions, the other poor and with 'bad' institutions: South and North Korea, the former West and East Germany, Israel and its neighbours, and the Dominican Republic and Haiti.

This answer undoubtedly contains much truth. No one considers it wrong, and many commentators consider it a full answer (and disagree with the studies discussed below). However, increasing numbers of economists find it incomplete, for two reasons. First, the answer merely notes national differences in institutional quality, and says nothing about



Figure 1 Rich and poor: skyscrapers and a shanty town, icons of contrasting conditions among the world's nations.

their origin. Why did Luxembourg, of all countries, end up with good institutions, whereas Nigeria didn't? Second, the answer neglects the non-institutional factors.

As regards origins, good institutions don't arise at random around the world.

Instead, they are outcomes of a long history shaped by geography, which helps to explain why Luxembourg has them but Nigeria doesn't. The origins of complex institutions are linked inextricably with the origins of states, which unfolded over thousands of years as by-products of sedentary, populous agricultural societies. These arose independently in only a few areas of the world endowed with many domesticable wild plant and animal species, beginning around 8500 BC in the Fertile Crescent of the Middle East¹⁵. In particular, state societies gradually evolved national loyalty instead of clan loyalty, deep experience of centralized government, pools of trained administrators and educated, literate citizens, and enforcement of social norms through government-administered laws rather than through individuals taking matters into their own hands. It proves difficult to telescope those developments of millennia into a generation through imitation and foreign aid.

Two studies^{1,2} by economists demonstrate this relevance of historical geography to institutional origins. In one study, Hibbs and Olsson¹ compared 112 nations' per capita wealth with the time since the local transition from hunting/gathering to agriculture. Two conclusions emerged: the larger the local biogeographic endowment of domesticable, large wild mammals and large-seeded wild cereals, the earlier was that transition locally; and the earlier that local transition, the richer is the country today. Part of the explanation is that some (but not all) countries with a long history of complex agricultural societies ended up with good institutions: geography and biogeography account for 40% of the explained variance in institutional quality.

In the other study², Bockstette *et al.* examined the growth rate of per capita wealth in the past 50 years, instead of current wealth itself. It turned out that countries with a long history of state societies have recently tended to enjoy high growth: countries that 50 years ago were still poor but had already developed complex institutions caught up quickly, once they added advanced technology to their institutional advantages.

For example, around 1950, when South Korea, Ghana and the Philippines were equally poor, most economists predicted that resource-rich Ghana and the Philippines were on the verge of wealth, whereas South Korea was doomed to remain mired in poverty. The result, of course, has been the opposite, because for 1,300 years South Korea has formed half of a unified, literate kingdom, and was strongly influenced by neighbouring China (one of the world's two oldest agricultural civilizations) long before that, whereas Ghana and the Philippines were exposed to rudimentary state government only within the past few centuries. As another example, Iceland, until a century ago Europe's poorest

country, is now among the world's ten richest despite its modest resources, while resource-rich Zambia is still poor. But Zambia acquired colonial state government barely a century ago, whereas Iceland has been a literate state for 1,100 years.

Besides those historical legacies, geography may also contribute to wealth through its effects on public health, agricultural productivity and transport costs^{3,4}. The first two of those factors penalize tropical countries, and the last penalizes land-locked countries.

As regards public health, tropical countries tend to carry much heavier disease burdens than do those of temperate zones, because parasites and disease vectors thrive all year round in the tropics but not in temperate climates. Disease is obviously bad for economies: workers who have spent years training have lower productivity and fewer years to contribute to the labour force than they otherwise would; high child mortality drives parents to bear many children in the hope that some will survive, so that frequent pregnancy or lactation makes women less able to join the work force; and health costs drain government budgets.

As for agricultural productivity, one might at first expect higher crop yields in year-round tropical climates than in temperate zones. But the reverse is true, for reasons that include parasites and crop pests.

Finally, transport costs by boat are lower than those over land — and hence lower for countries with a sea coast, or with big navigable rivers, than for land-locked countries without such rivers. That helps to keep countries such as Afghanistan, Bolivia, Chad, Laos, Mali and Zambia poor.

The burdens on health and agriculture explain why tropical countries are on average poorer than those of temperate regions, and why until recently the more tropical parts of the United States and Brazil were poorer than their temperate parts. Proof of the pudding comes from Southeast Asia's 'tiger' economies, which have achieved spectacular growth in the past half-century. Hong Kong, Malaysia, Mauritius (in the Indian Ocean), Singapore, Taiwan and Thailand became rich precisely by recognizing their tropical penalties, and by investing heavily in overcoming them through public-health measures, family planning, and developing economic sectors other than agriculture (Fig. 1).

Moving beyond geography, there are at least three non-geographic explanations for differing national wealth: the paradoxical curse of natural resources, reversals of fortune after colonization, and environmental damage. Taking first the resource curse, a common-sense prediction is for countries rich in natural resources, such as minerals, oil and tropical hardwoods, to be wealthier than countries not so endowed. In fact, countries deriving much of their income or foreign exchange from natural resources

— such as the Congo and Nigeria — are paradoxically poor^{5–7}. Among the suggested reasons for this are that dependence on natural resources promotes civil wars (with people of resource-rich provinces seceding to control their local resources); it creates temptations for corruption; and it raises prices and wages, thereby stunting the growth of manufacturing and other economic sectors.

Acemoglu *et al.*^{8,9} used the term “reversal of fortune” to explain differing economic changes among non-European countries colonized by European states in the past 500 years. The areas whose pre-colonial native societies had been richest (for example, Bolivia and Peru as parts of the Inca Empire, Mexico as the centre of the Aztec Empire, and India) are now poorer than pre-colonially poor areas such as Australia, Canada, New Zealand and the United States. Acemoglu *et al.* note, by way of interpretation, that areas that were formerly rich and densely populated but afflicted with tropical diseases were settled by few European colonists, who siphoned wealth from local people by exploitative institutions that today are bad for their economies as independent countries. In contrast, poorer areas where Europeans did not suffer high disease mortality did attract European settlers, who introduced institutions like those in their mother countries and more conducive to development.

Finally, a moment's reflection will suggest an objection to the idea that early agricultural origins lead to wealth^{1,2}. By that argument, the world's richest countries should now be those where agriculture arose earliest: Iraq, Iran and Syria. In fact, all three are poor

today, and would be even poorer were it not for oil. What happened? Their inhabitants had the misfortune to be living in fragile environments with low rainfall or high subsurface salt, which over millennia became damaged by deforestation, overgrazing, soil erosion and salinization. Those countries, too, experienced a reversal of fortune, but one due to environmental deterioration rather than colonial history¹⁶. Hence, a message from the studies described here is that aid donors should invest not only in institutions but also in public health, family planning and environmental protection. ■

Jared Diamond is in the Departments of Geography and Environmental Health Sciences, 1255 Bunche Hall, University of California, Los Angeles, California 90095-1524, USA.

e-mail: jdiamond@geog.ucla.edu

- Hibbs, D. A. Jr & Olsson, O. *Proc. Natl Acad. Sci. USA* **101**, 3715–3720 (2004).
- Bockstette, V., Chanda, A. & Putterman, L. *J. Econ. Growth* **7**, 347–369 (2002).
- Sachs, J. D. *et al. Sci. Am.* **284**, 62–67 (2001).
- Sachs, J. D. & Malaney, P. *Nature* **415**, 680–685 (2002).
- Sachs, J. D. & Warner, A. M. *J. Afr. Economies* **6**, 335–376 (1997).
- Sachs, J. D. & Warner, A. M. *Eur. Econ. Rev.* **45**, 827–838 (2001).
- Ross, M. *World Politics* **53**, 325–361 (2001).
- Acemoglu, D., Johnson, S. & Robinson, J. A. *Am. Econ. Rev.* **1**, 1369–1401 (2001).
- Acemoglu, D., Johnson, S. & Robinson, J. A. *Q. J. Econ.* **117**, 1231–1294 (2002).
- Smith, A. *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776; reprinted by Penguin, London, 1999).
- North, D. C. & Thomas, R. P. *The Rise of the Western World: A New Economic History* (Cambridge Univ. Press, 1973).
- Hall, R. E. & Jones, C. R. *Q. J. Econ.* **114**, 83–116 (1999).
- Knack, S. & Keefer, P. *Econ. Politics* **7**, 207–227 (1995).
- Easterly, R. & Levine, R. *J. Monet. Econ.* **50**, 3–39 (2003).
- Diamond, J. *Guns, Germs, and Steel: The Fates of Human Societies* (Norton, New York, 1997).
- Redman, C. O. *Human Impacts on Ancient Environments* (Univ. Arizona Press, Tucson, 1999).

Nanophysics

A step up to self-assembly

Kristen Fichthorn and Matthias Scheffler

Powerful computer simulations have resolved the mechanism for the nanoscale assembly of the 'hut'-like clusters that form after a few layers of atoms have been deposited on certain solid surfaces.

From the beautiful snowflakes that form from a random aggregation of water molecules, to the creation of a living organism, nature has found such efficient means of self-assembly that, in contrast, human techniques often seem crude. Even in our most impressive technologies for fabricating microstructures on surfaces — such as the lithographic techniques used to create integrated circuits — human efforts still seem like chiselling patterns out of stone. At the nanometre scale, the resolution that can be achieved using lithography is reaching its limit, and a new set of tools is needed. By better understanding nature's methods for assembly on solid surfaces, involving diffusion, nucleation and growth, it might be

possible to orchestrate these phenomena such that a complete computer chip consisting of several billion transistors could assemble itself, like a complex biological organism. Indeed, for nanotechnology to become affordable, nanostructures will have to build themselves; normal manufacturing methods will be useless. The laws of physics do not preclude this possibility, but our present understanding of surface physics is still too shallow to achieve such complex self-organization and assembly.

Zhu and colleagues¹, writing in *Physical Review Letters*, have quantified one of nature's mechanisms for creating nanometre-scale objects when atoms are deposited and grow into thin films on a solid surface of the same

material (called thin-film homoepitaxy). Using computer simulations and a first-principles approach², these authors model how atoms 'rain' down on to a solid surface and diffuse across it, by hopping between surface binding sites. The atoms may aggregate to form nuclei, which then grow into islands. Whether an island retains a two- or three-dimensional shape depends, in the conventional perspective, on how easily atoms can 'step down' from the top of the island to the lowest unfilled layer of the developing film.

The first atomic-level insight into this process came in 1966 from the elegant field-ion-microscopy studies led by Ehrlich³ and concurrent studies led by Schwoebel⁴. Ehrlich observed individual tungsten atoms hopping between binding sites on a tungsten surface and noted that when atoms approached downward step edges, they were often reflected away. This apparent repulsion was explained with bond-counting arguments: on flat crystallographic terraces, atoms have more bonds to other substrate atoms than they do if they are near the top of a step edge; there, the loss of substrate-atom bonds makes them less secure and hence they can be reflected from the edge. The energetic 'unwillingness' of atoms to descend step edges is now quantified by the Ehrlich-Schwoebel barrier.

On the basis of the same bond-counting arguments, homoepitaxial atoms residing at the bottom of a step edge of a close-packed surface should be the most secure, because they have a full complement of neighbouring atoms in the substrate below them and additional neighbours in the step edge beside them. It seems unlikely that these highly secure atoms would move upwards onto the tops of islands. However, evidence has been uncovered⁵ for such upward motion in experimental studies of aluminium homoepitaxy on its less close-packed (110) surface, using atomic force microscopy and low-energy electron diffraction. After about ten layers of atoms had been deposited, at a temperature of about 400 K, nanocrystalline 'huts' were seen to emerge rapidly from the substrate, growing to a height of about 50 nm (equivalent to more than 200 atomic layers) after the deposition of only another 20 layers (Fig. 1a). The emergence of these structures indicates significant upward motion of atoms.

To investigate this unusual occurrence, Zhu *et al.*¹ used density-functional theory, which allows the quantification, with the best accuracy available, of the many-electron nature of chemical bonding at surfaces and how it dictates energy barriers, mechanisms and timescales for atomic motion. They show that on the (110) surfaces of a variety of metals (including aluminium), atoms ascend steps by incorporating themselves into the step edge and pushing a step-edge atom onto the top of the step (Fig. 1b). Interestingly, the energy barriers for this upward

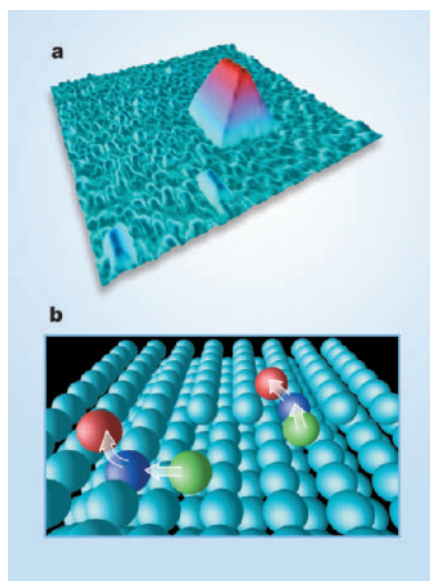


Figure 1 Step-climbing and hut-building. a, A nanocrystalline 'hut', formed by a cluster of aluminium atoms on an aluminium (110) surface². A total of 30 atomic layers have been deposited. The area of this atomic-force microscope image is $0.82 \times 0.82 \mu\text{m}^2$. b, Zhu *et al.*¹ suggest that hut-building is initiated by the upward diffusion of atoms at step edges on the (110) surface, even though this had previously been considered energetically unfavourable. An atom (green) at a step edge pushes an atom (blue) in the edge of the step onto the top (red), taking its place in the step.

motion can be lower than barriers for downward motion.

If atoms prefer to move upwards, why do they wait until ten layers have been deposited before beginning to form these monumental structures? To understand the relationships between the atomic processes that lead to assembly, Zhu *et al.* incorporated their results using density-functional theory into a simulation of the statistical mechanics of

the system, called an 'ab initio kinetic Monte Carlo' simulation^{2,6,7}. The process of thin-film growth is simulated as a series of discrete events by choosing and actuating various atomic-scale processes (such as atomic deposition, atom hopping on a terrace, down a step, up a step, and so on) with probabilities based on their timescales.

Their simulations verify experimental observations: a few rough layers act as necessary precursors from which nanocrystals arise. The islands develop into small mounds whose sides form 'minifacets', or small surfaces with a different crystallographic structure from that of the (110) substrate. In this case, atoms race up the sides of the minifacets even faster than they step up single steps (such as those shown in Fig. 1b) and actuate the rapid rise of the huts.

Zhu *et al.*¹ have unravelled one of nature's secrets for self-assembly. But there is still a long way to go to achieve true mastery of the art, such that magnetic memory devices, catalysts or integrated circuits can be fabricated simply by throwing atoms onto a surface and letting them organize themselves.

Kristen Fichthorn is in the Department of Chemical Engineering, Pennsylvania State University, University Park, Pennsylvania 16802, USA.

e-mail: fichthorn@psu.edu

Matthias Scheffler is at the Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4–6, 14195 Berlin-Dahlem, Germany.

e-mail: scheffler@fhi-berlin.mpg.de

1. Zhu, W., Buatier de Mongeot, F., Valbusa, U., Wang, E. G. & Zhang, Z. *Phys. Rev. Lett.* **92**, 106102 (2004).
2. Ruggerone, P., Ratsch, C. & Scheffler, M. in *Growth and Properties of Ultrathin Epitaxial Layers. The Chemical Physics of Solid Surfaces* Vol. 8 (eds King, D. A. & Woodruff, D. P.) 490–544 (Elsevier, Amsterdam, 1997).
3. Ehrlich, G. & Hudda, F. G. *J. Chem. Phys.* **44**, 1039–1049 (1966).
4. Schwoebel, R. & Shipsey, E. J. *J. Appl. Phys.* **37**, 3682–3686 (1966).
5. Buatier de Mongeot, F. *et al. Phys. Rev. Lett.* **91**, 016102 (2003).
6. Fichthorn, K. A., Merrick, M. L. & Scheffler, M. *Appl. Phys. A* **75**, 17–23 (2002).
7. Kratzer, P. & Scheffler, M. *Comput. Sci. Eng.* **3**, 16–25 (2001).

Molecular medicine

The writing is on the vessel wall

Christopher H. Contag and Michael H. Bachmann

Using an integrated approach, it is possible to identify the molecular addresses on blood-vessel walls that are read by blood cells — and, at least in rats, to deliver imaging agents and drugs to specific tissues.

Watching blood cells move through the blood vessels of living tissues under the microscope, researchers have learned that most of the time the cells move rapidly with the blood flow without stopping. Sometimes, however, they contact the walls of the vessel, slow down and roll along — perhaps even stopping and sticking in specific areas^{1–4}. This and other indirect evidence has suggested that the

endothelial cells that line the blood-vessel wall display different proteins on their surfaces according to where they are in the body and what events are occurring in a given organ. These proteins constitute the writing on the vessel walls — molecular 'addresses' that could tell circulating blood cells where they are and what to do in specific circumstances⁵.

It follows that scientists and physicians

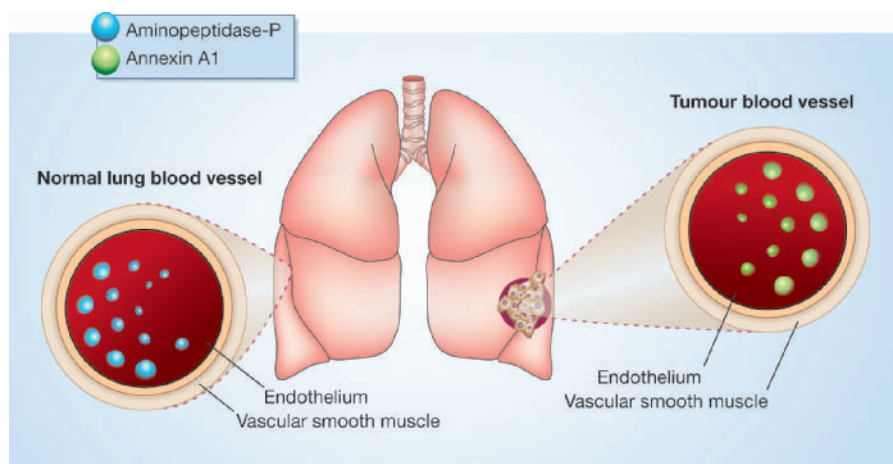


Figure 1 Molecular mapping. Oh *et al.*⁶ have used an approach involving proteomics, bioinformatics and molecular imaging to identify proteins — two of which are shown here — that are expressed exclusively on the blood-exposed surface of endothelial cells in the lungs and in lung tumours of rats. (Endothelial cells, surrounded by a sheath of vascular smooth muscle, make up the inner wall of blood vessels.) This approach might therefore provide a way of directing imaging and therapeutic compounds to specific healthy or diseased tissues.

might also be able to take advantage of these addresses, using them to deliver imaging or therapeutic agents to specific organs, to sites of tissue damage, or to a developing tumour. Towards this end, on page 629 of this issue Oh *et al.*⁶ describe an integrated approach — involving proteomics, bioinformatics and molecular imaging — by which to identify and characterize the proteins on the blood-exposed surfaces of endothelial cells in lung and lung-tumour tissues. The authors also demonstrate that these proteins can be effectively targeted for imaging and therapy.

A number of proteins, for example members of the integrin family, have already been found to be expressed differentially on the blood-vessel walls of different tissues, and have been identified as promising targets to which to selectively deliver molecules for tumour imaging and treatment^{7–9}. Nevertheless, researchers would benefit greatly from a complete map of the proteins on the blood-exposed surfaces of endothelial cells. A major problem is that these surfaces represent only a small fraction of any given tissue — so the proteins presented there are the proverbial needles in the haystack. The identification of subtle or under-represented blood-vessel markers therefore requires a focused approach to reduce the overall complexity of tissue proteins and yet to preserve the diversity of markers in the blood-vessel wall.

Oh and co-workers^{6,10} have come up with a clever way of doing just that. Analogous to using a magnet to extract needles from a haystack, these authors used a rapid, affinity-based isolation procedure to enrich and purify, from organs including rat lung and lung tumours, the parts of blood-vessel endothelial cells that contact the blood. They accomplished this by infusing colloidal silica particles into the bloodstream of rats, where these particles attached to the endothelial cells. Subsequent centrifugation

of tissue homogenates allowed endothelial-cell membranes and attached organelles called caveolae¹⁰ to be separated from the remainder of the cells. Caveolae are small invaginations of the cell surface that carry out transport and signalling functions in cells. They are lined with an indicator protein called caveolin as well as other cell-surface proteins. For the final purification step, an antibody that recognizes caveolin, coupled to magnetic beads, was used to isolate caveolae and their associated proteins.

In the current work⁶, the authors analysed the components of this enriched cellular subset with mass spectrometry, database searches and antibodies for proteins that are found only in the wall of the blood vessels of given tissues, and then focused on the lungs and lung tumours. Next, they subjected those proteins that they had found exclusively in a given tissue vasculature to structural analyses, to ascertain which were most probably present on the endothelial cell surfaces and exposed to the circulation. Out of potentially hundreds of thousands of cellular proteins, Oh *et al.* extracted 37 that were present only in the endothelial membrane; 11 of these were probably in a configuration that includes an extracellular portion that could be presented to blood cells.

To validate this approach in an animal model, the authors used radiolabelled antibodies that recognize 2 of the 11 identified proteins — aminopeptidase-P and annexin A1 — to show that these proteins are found specifically in the blood vessels of lungs and lung tumours, respectively, in rats (Fig. 1). This brought the study full circle, from intact organs, to proteomic and bioinformatic analyses of selected proteins, to the imaging of specific signatures on the walls of blood vessels in defined organs of a living subject. As a *coup de grâce*, Oh *et al.* found that the radiolabelled antibodies against annexin A1

also prolonged survival, and may have caused significant remission of solid lung tumours in rats.

So, by learning what blood cells already know, we might be able to use the writing on the vessel wall both to design unique molecular tools by which to recognize specific diseases, and to direct therapies to specific organs⁹ — thereby avoiding sending drugs as unwanted ‘spam’ to all other organs in the body. For instance, proteins in the vasculature of tumours could enable imaging agents to locate growing malignancies, and drugs to specifically inhibit the tumour’s endothelial cells and so prevent further tumour growth⁹.

It is likely that blood vessels in diseases other than cancer will have unique signatures as well; for example, the proteins expressed in and around atherosclerotic plaques might be used to identify (by imaging) and treat certain cardiovascular diseases. In many instances, therapies will need to be delivered to tissue cells beyond the blood-vessel wall. But Oh and colleagues’ strategy might actually provide a solution to this problem, too. Given that the molecular targets they identified exist in the caveolae of endothelial cells, perhaps the role of these organelles in taking up molecules could be exploited to transport drugs across the endothelial barrier to reach cells beyond the endothelium.

To take this to the next level, perhaps it might even be possible to rebuild damaged tissues within the body through this type of tissue targeting. Imagine directing appropriate stem cells, known to promote tissue repair¹¹, to sites of tissue damage in a patient. Despite the current controversy over the potential of adult stem cells from one tissue to form other types of cells (see, for example, refs 12, 13), advances in understanding stem-cell biology, cell-to-cell signalling and tissue remodelling, coupled with effective tissue targeting, might provide the necessary tools. If so, the writing would indeed be on the wall for all sorts of diseases.

Christopher H. Contag is in the Departments of Pediatrics, Microbiology & Immunology, and Radiology, and Michael H. Bachmann is in the Department of Pediatrics, Stanford University School of Medicine, Stanford, California 94305, USA. e-mails: ccontag@cmgm.stanford.edu
bachmann@cmgm.stanford.edu

1. Miller, M. J., Wei, S. H., Cahalan, M. D. & Parker, I. *Proc. Natl. Acad. Sci. USA* **100**, 2604–2609 (2003).
2. M’Rini, C. *et al. J. Exp. Med.* **198**, 1301–1312 (2003).
3. Condeelis, J. & Segall, J. E. *Nature Rev. Cancer* **3**, 921–930 (2003).
4. Cahalan, M. D., Parker, I., Wei, S. H. & Miller, M. J. *Curr. Opin. Immunol.* **15**, 372–377 (2003).
5. Folkman, J. *Cancer Cell* **1**, 113–115 (2002).
6. Oh, P. *et al. Nature* **429**, 629–635 (2004).
7. Pasqualini, R. & Ruoslahti, E. *Nature* **380**, 364–366 (1996).
8. Pasqualini, R., Arap, W. & McDonald, D. M. *Trends Mol. Med.* **8**, 563–571 (2002).
9. Hood, J. D. *et al. Science* **296**, 2404–2407 (2002).
10. Oh, P. & Schnitzer, J. E. *J. Biol. Chem.* **274**, 23144–23154 (1999).
11. Cowan, C. M. *et al. Nature Biotechnol.* **22**, 560–567 (2004).
12. Murry, C. E. *et al. Nature* **428**, 664–668 (2004).
13. Balsam, L. B. *et al. Nature* **428**, 668–673 (2004).

Signal transduction

Ras finds itself a sticky spot

Cell **117**, 637–648 (2004)

The Ras family of proteins, which regulate cell proliferation and are mutated in around 30% of human cancers, were thought to act mainly in the cellular outer membrane. But Andrew K. Sobering *et al.* have found that, in yeast at least, a Ras protein also works in the intracellular compartment known as the endoplasmic reticulum (ER).

Previously, the authors identified a yeast protein called Eri1 that interacts with Ras and which, when mutated, has similar effects to Ras alterations. Sobering *et al.* have now found that Eri1 is part of an enzyme complex in the ER that is involved in producing glycosylphosphatidylinositol (GPI) anchors. These anchors are attached to proteins destined for the cell wall and help yeast cells to alter their shape and stick to extracellular surfaces.

The authors go on to show that yeast Ras binds and blocks this enzyme complex, so cutting the production of GPI anchors. Should the same prove true in humans, this suggests an additional way in which Ras mutations might help cancers to spread — by lowering the amounts of GPI-anchored proteins on the surface of tumour cells, so allowing them to better invade surrounding tissue or spread to other organs.

Helen Pearson

Materials science

Self-cleaning clothes

J. Am. Ceram. Soc. **87**, 953–955 (2004)

In the classic 1951 film *The Man in the White Suit* (pictured), Alec Guinness played a scientist who invents a fabric that never gets dirty or wears out. A chemists' pipe-dream perhaps, but the prospect of a self-cleaning fibre might now be a step closer.

Walid A. Daoud and John H. Xin have found an efficient way of coating cotton cloth with nanoparticles of titanium dioxide. These particles are catalysts that help to break down organic materials, requiring only sunlight to trigger the reaction. For maximum activity, the nanoparticles must have the correct 'anatase' crystal structure, which has previously been difficult to achieve in these tiny grains.

The authors dipped a small cotton patch into a suspension of titanium dioxide held in a mixture of water, ethanol and acetic acid at about 40 °C. After just half a minute the cloth was removed, padded dry and heated to 97 °C in an oven for 15 minutes. Three hours in boiling water completed the process. The authors report a good covering of anatase nanoparticles, each



of which measures about 20 nm across.

They speculate that catalyst-coated materials could one day lead to self-cleaning fabrics that tackle organic dirt, environmental pollutants and harmful microorganisms. But they should perhaps take note of the film's plot: fearing that his invention might threaten their livelihoods, textile manufacturers took Guinness prisoner to keep his remarkable fabric under wraps.

Mark Peplow

Plant development

Pollen redirection

Development **131**, 2707–2714 (2004)

Unlike animals, plants do not have motile sperm. Instead, pollen, which carries the male gamete, grows down through the plant's female organs in search of ovules to fertilize. Ana Margarida Prado and colleagues now provide evidence that nitric oxide (NO) may be involved in directing pollen to its goal.

The authors found that, when a source of NO was placed close to the tip of a growing pollen tube, the tube executed a turn of more than 90° to grow away from the gaseous signalling molecule. This behaviour could be abolished by adding an NO scavenger to the growth medium, but was enhanced by adding sildenafil citrate (Viagra), showing that cyclic GMP molecules — the breakdown of which is inhibited by Viagra — might be involved in transducing the NO signal in pollen tubes.

The authors also found that pollen tubes themselves synthesize NO, in organelles positioned behind the growing tip; the resulting NO gradient would maintain growth in a straight line. But pollen must make sharp turns to reach the ovary. The authors suggest that NO could also be produced at 'turning points' within the flower, to point pollen in the right direction.

Christopher Surridge

Neurobiology

Sense and sensitivity

Neuron **42**, 595–605 (2004)

Rod cells in the retina can detect very small changes in illumination. But how do they accurately transmit these changes to downstream nerve cells? The secret, according to Wallace B. Thoreson *et al.*, seems to lie in their unusual response to intracellular calcium, the concentrations of which alter in response to light.

Using rods from the eyes of tiger salamanders, the authors manipulated the cells to release calcium from caged calcium compounds in a controlled way. They then measured the cells' responses by analysing the electrical properties of the cell membranes to determine how many neurotransmitter molecules were being released.

In other nerve cells, the relationship between calcium concentration and neurotransmitter release is nonlinear. This is because several calcium ions must bind to a neurotransmitter-containing vesicle before the vesicle can fuse with the cell membrane and release the neurotransmitters. But Thoreson *et al.* find that rods contain a highly calcium-sensitive pool of vesicles that, over the range of calcium levels thought to be present normally, show a linear relationship between calcium concentration and neurotransmitter release. This allows rods to relay information about small changes in illumination with high responsiveness.

Laura Nelson

Applied physics

Reinventing the light bulb

Appl. Phys. Lett. **84**, 4869–4871 (2004)

Thomas Edison's first incandescent light bulbs had filaments made from carbon (carbonized cotton thread), but these were eventually replaced with longer-lasting tungsten. Now carbon filaments might make a comeback, thanks to the discovery by Jinquan Wei and colleagues that bundles of carbon nanotubes provide robust filaments in household bulbs and have some advantages over tungsten.

The researchers replaced the filaments of standard safelights (used in photographic darkrooms) with strands composed of close-packed single- and double-walled nanotubes. These carbon filaments switch on at lower voltages (3–5 V) than tungsten, glow more brightly at the same voltage, and use less power. They survive more than 5,000 on–off cycles and operate continuously for over two weeks without degradation. The glow seems to result from a combination of ohmic heating and, at higher voltages, electroluminescence from the nanotubes.

Phillip Ball

Pterosaur embryo from the Early Cretaceous

An impressive fossil discovered in China confirms that pterosaurs were egg-layers.

Dinosaur embryos have been discovered all over the world^{1,2}, but so far no pterosaur embryos have been reported. Here we describe a Chinese fossil from the Early Cretaceous period containing an embryo that is unambiguously a pterosaur. The embryonic skeleton, which is exquisitely preserved in its egg, is associated with eggshell fragments, wing membranes and skin imprints. This discovery confirms that pterosaurs were egg-layers and sheds new light on our understanding of pterosaur development.

The fossil was collected from the lacustrine shales of the Jingangshan Bed of the upper Yixian Formation at Jingangshan in western Liaoning, China. ⁴⁰Ar/³⁹Ar dating³ shows the fossil-bearing deposit is 121 million years old, indicating that the fossil is from the late Early Cretaceous (Aptian) period. Associated fossils from the same horizon include *Lycoptera*, *Manchurochelys*, *Yabeinosaurus* and some undescribed pterodactyls and birds⁴.

The embryo is preserved in the brown part and counterpart of an almost complete egg (Fig. 1a, b). There is a distinctive elliptical margin that is markedly different from the yellowish-grey matrix. The egg has a maximum length of 53 mm and a maximum width of 41 mm.

The skull is preserved in ventral view, whereas the postcranial bones are in dorsal view. The skull is curved backwards, forming an inverse U-shape with the vertebral column, which extends along the long axis of the egg. The forelimbs are tightly folded. Both the lower jaw and the limb bones are characterized by less ossified epiphyses and an immature grainy texture⁵. The scapula is not fused with the coracoid. The vertebral column, carpal bones and pedal digits are also incompletely ossified.

The skeleton has a partial skull and nearly complete postcranial bones (Fig. 1a–d). The lower jaws are robust, with two slender and slightly curved teeth. At least five dorsal vertebrae are preserved. The humerus is strong, with a non-rectangular deltopectoral crest. The wing metacarpal is well developed, and is slightly shorter than the humerus. The length of the femur nearly equals the length of the humerus, and is about three-quarters of the length of the tibia. Metatarsals I–IV are relatively small and slender; metatarsal III is short and about 20% of the length of the tibia.

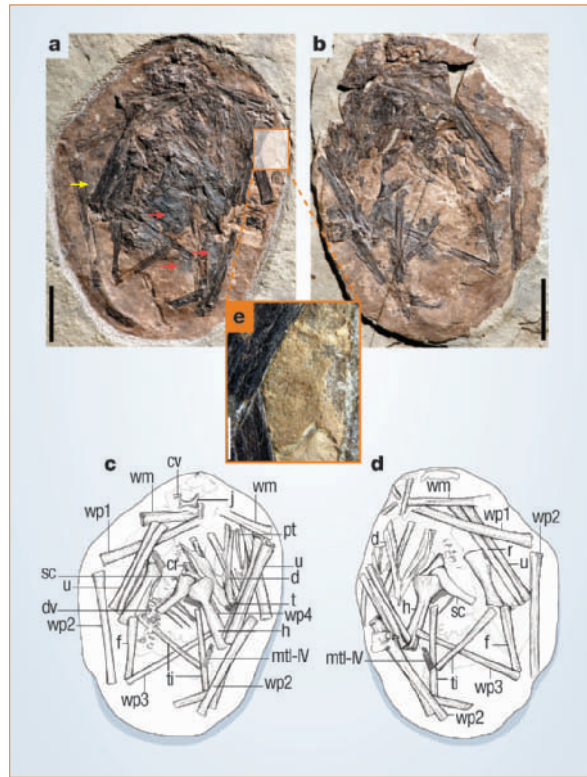


Figure 1 Pterosaur embryo inside an egg from the Early Cretaceous period from Liaoning, China (IVPP V13758). **a–d**, Photographs of part (**a**) and counterpart (**b**) of the fossil and their corresponding line drawings (**c**, **d**; not to scale). Red arrows indicate skin imprints and the yellow arrow indicates the fibres of the wing membrane. Scale bar, 10 mm. **e**, Close-up of the papilla-like ornamentation of the eggshell (corresponding to orange frame in **a**). Scale bar, 2 mm. Abbreviations in **c**, **d**: cr, coracoid; cv, cervical vertebra; d, dentary; dv, dorsal vertebra; f, femur; h, humerus; j, jugal; mtl–IV, metatarsals I–IV; pt, pteroid; r, radius; sc, scapula; t, tooth; ti, tibia; u, ulna; wp1–4, first to fourth phalanges of the wing digit; wm, wing metacarpal.

The embryo can undoubtedly be identified as that of a pterosaur because the preserved skeleton shows an extremely elongated fourth finger and a stout humerus with a well developed deltopectoral process⁶. The remarkably elongated wing metacarpal indicates that it is a pterodactyl⁷.

We could assign the embryo to the Ornithocheiridae because of such features as its relatively short metatarsals (less than 25% of the length of the humerus) and the rounded, flange-shaped deltopectoral crest of the humerus^{8,9}. Of the pterosaurs that have been discovered from the Yixian Formation^{10,11}, *Haplopterus* most resembles this embryo. However, allometric growth during pterosaur development means we must be cautious about making direct ratio comparisons between embryonic and adult elements, so it is uncertain whether this fossil can be assigned to any known taxa of the Ornithocheiridae.

Preserved patches of eggshell showing papilla-like ornamentation (Fig. 1e) are

extensively distributed in the fossil, and may assist identification of the embryo. Wing membrane fibres are also preserved, together with some wing phalanges. Large patches of skin imprints are preserved, mainly in the posterior part of the body. Preservation of such delicate tissues with the skeleton and eggshell probably indicates that the embryo was killed and deposited quickly as a result of a natural disaster, such as a volcanic eruption⁴.

This embryonic skeleton from the Jehol Biota is larger than the smallest baby *Pterodactylus*, from Solnhofen, which had proper flight capability and a wingspan of about 18 cm (ref. 7). The Liaoning embryo has a wingspan of 27 cm, indicating that the embryo would have grown up into a medium-to-large pterosaur.

The Liaoning fossil has preserved nearly all the skeletal elements and clear imprints of the wing membranes, indicating that this pterosaur embryo was probably enjoying its last few days in the egg before hatching out to walk on the Early Cretaceous earth.

Xiaolin Wang, Zhonghe Zhou
Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, PO Box 643, Beijing 100044, China

e-mail: xlinwang@263.net

1. Norell, M. A. *et al.* *Science* **266**, 779–782 (1994).
2. Chiappe, L. M. *et al.* *Nature* **396**, 258–261 (1998).
3. Smith, P. E. *et al.* *Can. J. Earth Sci.* **32**, 1426–1431 (1995).
4. Wang, X. & Zhou, Z. in *The Jehol Biota* (eds Chang, M., Chen, P., Wang, Y.-Q. & Wang, Y.) 19–36 (Shanghai Scientific & Technical, Shanghai, 2003).
5. Bennett, S. C. *Paleobiology* **19**, 92–106 (1993).
6. Unwin, D. M. in *Evolution and Palaeobiology of Pterosaurs* (eds Buffetaut, E. & Mazin, J. M.) *Geol. Soc. Lond., Sp. Publ.* **217**, 139–190 (2003).
7. Wellnhofer, P. *The Illustrated Encyclopedia of Pterosaurs* 1–192 (Salamander, London, 1991).
8. Unwin, D. M. *Mitt. Mus. Naturk., Berl., Geowiss. Reihe* **4**, 189–221 (2001).
9. Unwin, D. M., Lü, J. & Bakhrina, N. N. *Mitt. Mus. Naturk., Berl., Geowiss. Reihe* **3**, 181–206 (2000).
10. Ji, S., Ji, Q. & Padian, K. *Nature* **398**, 573–574 (1999).
11. Wang, X. & Lü, J. *Chinese Sci. Bull.* **46**, 1112–1117 (2001).

Competing financial interests: declared none.

brief communications arising online
www.nature.com/bca

Palaeobiology: Dutch diaries and the demise of the dodo

J. P. Hume, D. M. Martill & C. Dewdney
 (doi:10.1038/nature02688)

Palaeobiology

Dutch diaries and the demise of the dodo

Arising from: D. L. Roberts & A. R. Solow *Nature* **426**, 245 (2003).

Roberts and Solow¹ have presented theoretical data to suggest that the dodo, *Raphus cucullatus*, of Mauritius may have persisted until 1690, some 28 years beyond its accepted extinction date of 1662. Here we analyse new historical data derived from records of hunting caches, noted by Oppershoofd (Chief) Isaac Joan Lamotius, which confirm that specimens of *R. cucullatus* were collected regularly for at least 26 years beyond 1662. We calculate a new extinction date, which differs by only three years from that calculated by Roberts and Solow, but which, using a statistical treatment similar to theirs, greatly narrows the confidence interval.

The last accepted record of the dodo is that from the shipwrecked mariner Volkert Evertszoon, who in 1662 walked on to an islet that supported a small population of dodos². In 1674, however, Commandeur Hubert Hugo questioned a recaptured slave called Simon, who had seen dodos on two occasions³ during his time at large, between 1663 and 1674. Previously unreported is a note⁴ that hunters captured and killed dodo and other endemic prey for Hugo on 16 August 1673.

Following Hugo, Lamotius became Oppershoofd of Mauritius from 1677 to 1692 and he recorded events on the island in a series of diaries. During the first few years of his command, Lamotius was supplied with ample provisions from the Cape (South Africa) and so did not regularly send out hunting parties. But diaries from 1685 to 1688, which are held in the Cape Archive, suggest that Mauritius received little support during those years from either the Cape or the Netherlands⁵. Over that period, Lamotius dispatched hunters daily to obtain food, and he maintained a day-by-day account of their quarry⁶.

Although the importance of this document has been realized for some time⁷, its reliability has been questioned². Consequently, Roberts and Solow¹ excluded these records in their analysis of the dodo's demise. Lamotius records dodo (as "dodaersen") on 12 separate occasions as part of the hunters' quarry during the period 1685–88 (ref. 6).

As endemic species became rarer, the quarry mainly comprised animals that had been introduced to the island such as deer, pigs and goats. Ducks, geese and flamingos were abundant at certain times of the year, but only very rarely did the catch include the now extinct Mauritian giant tortoise (*Cylindraspis* spp.) or the dodo. Lamotius last recorded the capture of a dodo on 25 November 1688.

A problem raised over Lamotius's

accounts is his use of the term dodaersen. Cheke² maintains that this name was transferred from the dodo to another flightless endemic Mauritius bird, the red rail (*Aphanapteryx bonasia*), which was much smaller and had a slender beak — quite different from the massive bill of the dodo.

We consider it much more likely that Lamotius was using the name dodaersen to refer to the dodo. He was a credible observer, having been instructed by the Dutch East



The dodo of Mauritius has long captured the imaginations of artists. Author Julian Pender Hume depicts the bird as it was before its extinction, which evidence now suggests occurred in 1693.

India Company to record, among other things, the natural history of Mauritius. He compiled a list of endemic trees and their medicinal uses on the island, and sent botanical specimens to the gardens of the company at the Cape; he was also an inventor and draughtsman³. Lamotius would have been well acquainted with the distinctive morphology of the dodo because of its fame in western Europe — at least three specimens had been transported to the Netherlands and many paintings were in existence, predominantly by Dutch artists.

For these reasons, it is unlikely that Lamotius would have confused the red rail with a dodo. The Dutch were familiar with rails and rail-like birds in their home territory, calling them velt-hoenders, meerkoe and waterhoen. There is no mention of these in the accounts of either Hugo or Lamotius, whereas earlier Dutch accounts do refer to the red rails of Mauritius as velt-hoenders (field chickens)². There is no evidence that Lamotius or Hugo changed this name to dodaersen.

The dates derived from the dodo records of Lamotius and the new records of Hugo confirm that the dodo was present on

Mauritius at a time close to the estimated extinction time of 1690 that Roberts and Solow calculate¹. This date has a 95% confidence interval of 1669 to 1797, calculated from their record of reported observations covering 1598–1662. Using the same statistical analysis¹ and all the observations together, including the new dates of Hugo and the records of Lamotius (up to 1688), we estimated the likely duration of isolated populations of dodos beyond the last records of Lamotius. We calculate a new estimated extinction date of 1693, with a 95% confidence interval of 1688 to 1715, a period of only 27 years. Our data set comprises 16 observations, and the last four occur within a four-year period.

Although our data have extended the estimated extinction time by only three years, our 95% confidence interval is considerably narrower (27 years as opposed to 128 years). This is due in part to a change in methods of data collection in that period: records seem to be *ad hoc* before 1685, but then observations become more frequent because of the need to hunt food.

Lamotius's hunting record ceased in 1688 because he became despondent, it is believed⁷; a diary for 1689 cannot be located; and Lamotius was arrested and left the island in 1692. Although the Dutch have been blamed for the final extinction of the dodo, we note that the time interval from 1688 to the time of the French takeover of Mauritius in 1710 (ref. 7) lies within the 90% confidence interval for the estimated extinction time. There is thus a 10% chance that the dodo became extinct during the French occupation of Mauritius.

Julian Pender Hume*†, David M. Martill*, Christopher Dewdney*

*Palaeobiology Research Group, School of Earth and Environmental Sciences, University of Portsmouth, Portsmouth PO1 3QL, UK

e-mail: j.hume@bun.com

†The Bird Group, Department of Zoology, Natural History Museum, Tring HP23 6AP, UK

doi: 10.1038/nature02688

1. Roberts, D. L. & Solow, A. R. *Nature* **426**, 245 (2003).

2. Cheke, A. S. *Studies of Mascarene Island Birds* (ed. Diamond, A. W.) 5–89 (Cambridge University Press, Cambridge, 1987).

3. Pitot, A. T'Eylandt Mauritius. *Esquisses Historiques (1598–1710). Précédés d'une notice sur la découverte des Mascareignes et suivies d'une monographie du dodo, des solitaires de Rodrigue et de Bourbon et de l'oiseau bleu* (Port Louis, Mauritius, 1905).

4. Hugo, H. Letter from Hugo to the Cape Commander. *Archive of the Political Council*, C293 (Cape Archives, South Africa, 1674).

5. Sleight, D. *The Economy of Mauritius during the Second Dutch Occupation (1664–1710) in Globalization and the Southwest Indian Ocean: Mauritius and Neighbouring Islands* (Proc. Int. Sem. 21–23 September 1998).

6. Lamotius, I. J. J. Dag-register van't principale gepasseerde op't Eylandt Mauritius gehou'den byj oppershoofd Isaac Joan Lamotius gebinnende mit den 1st January 1685–1688. *Archive of the Political Council*, C2238, C2239, C2240 (Cape Archive, South Africa, 1685–1688).

7. Moree, P. A. *Concise History of Dutch Mauritius, 1598–1710: A Fruitful and Healthy Land* (Kegen Paul, London and New York, 1998).

Eight glacial cycles from an Antarctic ice core

EPICA community members*

*A full list of authors appears at the end of the paper

The Antarctic Vostok ice core provided compelling evidence of the nature of climate, and of climate feedbacks, over the past 420,000 years. Marine records suggest that the amplitude of climate variability was smaller before that time, but such records are often poorly resolved. Moreover, it is not possible to infer the abundance of greenhouse gases in the atmosphere from marine records. Here we report the recovery of a deep ice core from Dome C, Antarctica, that provides a climate record for the past 740,000 years. For the four most recent glacial cycles, the data agree well with the record from Vostok. The earlier period, between 740,000 and 430,000 years ago, was characterized by less pronounced warmth in interglacial periods in Antarctica, but a higher proportion of each cycle was spent in the warm mode. The transition from glacial to interglacial conditions about 430,000 years ago (Termination V) resembles the transition into the present interglacial period in terms of the magnitude of change in temperatures and greenhouse gases, but there are significant differences in the patterns of change. The interglacial stage following Termination V was exceptionally long—28,000 years compared to, for example, the 12,000 years recorded so far in the present interglacial period. Given the similarities between this earlier warm period and today, our results may imply that without human intervention, a climate similar to the present one would extend well into the future.

The climate of the last 500,000 years (500 kyr) was characterized by extremely strong 100-kyr cyclicity, as seen particularly in ice-core¹ and marine-sediment^{2,3} records. During the earlier part of the Quaternary (before 1 million years ago; 1 Myr BP), cycles of 41 kyr dominated. The period in between shows intermediate behaviour, with marine records showing both frequencies and a lower amplitude of the climate signal^{2,3}. The observed frequencies arise from parameters of the Earth's orbit that control the amount, and the seasonal and latitudinal distribution, of solar radiation⁴. However, the reasons for the dominance of the 100-kyr (eccentricity) over the 41-kyr (obliquity) band in the later part of the record, and the amplifiers that allow small changes in radiation to cause large changes in global climate, are not well understood. New records of the earlier periods, looking at parameters unavailable in marine records, are needed.

Ice cores provide the most direct and highly resolved records of (especially) atmospheric parameters over these timescales. They record climate signals, as well as forcing factors of global significance such as greenhouse gases and of more regional significance such as atmospheric aerosol content. Until now, ice-core data have been available only for the past 420 kyr, with the longest record coming from Vostok in East Antarctica¹, supported by the 340-kyr record from Dome Fuji⁵. These data indicated the similarities of the last four glacial terminations. They showed that glacials and interglacials had similar bounds in the measured properties over the last four cycles. Most tellingly, they showed the very close association between greenhouse gases^{1,6} (CO_2 , CH_4) and climate (as recorded using the Antarctic temperature proxy, the deuterium/hydrogen ratio in ice, represented as δD) over this period. The Vostok record has become a compelling target against which other records and modelling efforts are tested.

The European Project for Ice Coring in Antarctica (EPICA) is a consortium of laboratories and Antarctic logistics operators from ten nations, with the goal of obtaining two deep ice cores in East Antarctica. The study of one core, from Kohnen Station in the Dronning Maud Land sector of Antarctica (see Supplementary Fig. 1) is aimed at producing a high-resolution record of at least one glacial–interglacial cycle in the sector of Antarctica facing the Atlantic Ocean, for comparison with Greenland records⁷. The second core (named EDC) from Dome C ($75^\circ 06' \text{S}$, $123^\circ 21' \text{E}$,

altitude 3,233 m above sea level), discussed here, is aimed at producing a record of the longest time period possible. The site⁸ has an ice thickness of $3,309 \pm 22 \text{ m}$; the current drilling depth is 3,190 m, of which 3,139 m has been analysed for a wide range of constituents. The current mean annual surface temperature is -54.5°C , and the snow accumulation rate is $25 \text{ kg m}^{-2} \text{ yr}^{-1}$ (2.5 cm water equivalent per year). The drill site is 56 km from the site of a previous Dome C core⁹ that provided records extending into the last glacial period, and 560 km from the site of the Vostok cores¹. The completion of the Dome C core was delayed when the first drilling became stuck at 788 m in 1999, and this shorter EDC96 core has already yielded many important results from the last 45 kyr (see, for example, refs 10–14).

Here we present the EDC records of δD and other parameters, analysed at low resolution, for the available core. We show that the core represents 740 kyr, including all of marine isotope stage (MIS) 11, which was not completed in the Vostok record, and running through a further three complete 100-kyr cycles, to MIS 18.4. We compare the amplitude and frequency structure of the period before MIS 11 with that of the more recent period. We focus in more detail, with new greenhouse-gas and ice-chemical data, on Termination V, from MIS 12 to MIS 11, discussing first the integrity of the record. The different parameters measured on this termination are then discussed in terms of similarities to and differences from younger terminations.

Stratigraphy of the EDC core

The ice-core data (see Methods) are reported in Fig. 1 as a function of depth. In this section, conductivity, grain size, dust and δD data, taken together, allow us to define a reliable stratigraphy of the core in terms of terminations and of broad correspondence with the deep-sea record. In the following section, we derive a timescale—which should be considered preliminary—and develop arguments supporting our claim that the core stratigraphy is undisturbed at the current depth (3,139 m) despite the relative proximity of the bedrock (less than 200 m).

Under the conditions at Dome C, both measurements (see Methods) of electrical conductivity¹⁵ are dominated by variations in the acidity of the ice¹⁶. This property does not vary in a simple way with climate, increasing in both very cold and very warm stages,

with the lowest values in intermediate climates. Cold periods in Antarctica are characterized by much greater dust fallout than is found during interglacials (for example, the Last Glacial Maximum (LGM)/Holocene ratio of 26 for dust flux¹³), related to a combination of increased aridity and wind strength. Large numbers of dust particles within the ice lead to a decrease in the ice-grain growth rate¹⁷. Consequently, each significant decrease of the average grain radius (Fig. 1) also marks an interglacial to glacial transition. The isotopic composition of the ice, δD (used here) and $\delta^{18}O$, is classically used as an indicator of temperature change. Isotopic models predict that δ values should vary linearly with temperature in mid- and high latitudes. There is now a series of arguments supporting the use of this present-day temperature/isotope spatial slope to interpret isotopic records from Antarctica^{18,19}, at least for deep ice cores from the East Antarctic plateau.

Electrical, dust and δD (Fig. 2) data can easily be matched between the EDC and Vostok cores into stage 11. We deduce that ice from 3,310 m at Vostok and from ~2,770 m at EDC corresponds to the same time period (423 kyr BP in the GT4 Vostok chronology). Transition V is then very clearly marked both in the dust, grain size and δD records with the coldest part of MIS 12 at around ~2,790 m (Fig. 1), and with Termination V (that is, the MIS 12 to MIS 11 transition) roughly corresponding to the depth interval between 2,790 and 2,760 m.

Below the dielectric-profiling peak corresponding to MIS 11 there is a large depth interval with low dielectric-profiling values. There is, however, a clear dust peak, as well as a large decrease in the average grain size, at a depth of 2,910 m, which should correspond to the cold MIS 14, thus implying that there is no dielectric-profiling peak within MIS 13. The δD record confirms that the interglacial MIS 13 peaks at a depth of 2,842 m, but is considerably colder than subsequent interglacials. This intermediate climate is insufficient to give a dielectric-profiling peak, probably because of reduced preservation of volatile acids²⁰.

From the δD record, we first note a clear change in the amplitude of glacial–interglacial changes before and after MIS 12, with the

older period being characterized by just one minimum as deep as those observed during the last 400 kyr, and by consistently lower maxima (by about 20‰). As discussed below, this change of amplitude corresponds to the mid-Brunhes climate shift (and does not result from some smoothing process in the ice). There is an excellent correspondence between the δD and the dust record and based on these we can assign the base of the transition at ~3,042 m to the next cold stage, MIS 16.2. In the deep-sea core record, stage 16.2 corresponds to particularly low sea level and was probably very cold. This is exactly what is seen in the δD where, before MIS 12, only stage 16 reached δD levels as low as those of the LGM. The next δD peaks (low dust) can then be attributed to full interglacial 17 and interstadial 18.3 with the bottom of the record corresponding to MIS 18.4.

Timescale and integrity of the deep ice

The timescale (called EDC2; see Methods) developed for the Dome C deep ice core is based on an inverse dating method²¹, constrained by a small number of control age windows, which are mainly set to glacial terminations by comparison to the marine records. The fact that a simple one-dimensional model with only four free parameters can be matched (to 3,139 m depth) so well, in both timing and shape, with the orbitally tuned marine records (Fig. 2c) is evidence for the integrity of the stratigraphy of the Dome C record. The good match extends through the period from 338 to 626 kyr, in which there are no imposed control windows. The difference between the ages of gas bubbles and the surrounding ice was computed with a firn model²².

The first section of EDC ice that is novel, that is, older than was

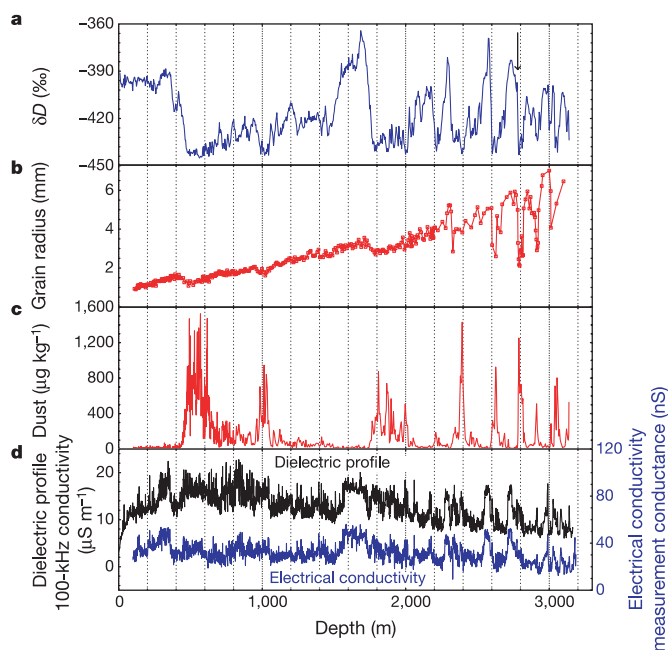


Figure 1 Measured parameters from the EPICA Dome C ice core, on an ice depth scale. **a**, δD , averaged over 3.85-m sections. **b**, Grain radius, measured approximately every 10 m. **c**, Dust concentration—below 787 m, there is one sample every 5.5 m; above that, one sample every 1.5 m. **d**, Electrical data (as discussed in the Methods), in 1-m averages. Termination V is marked by an arrow in **a**.

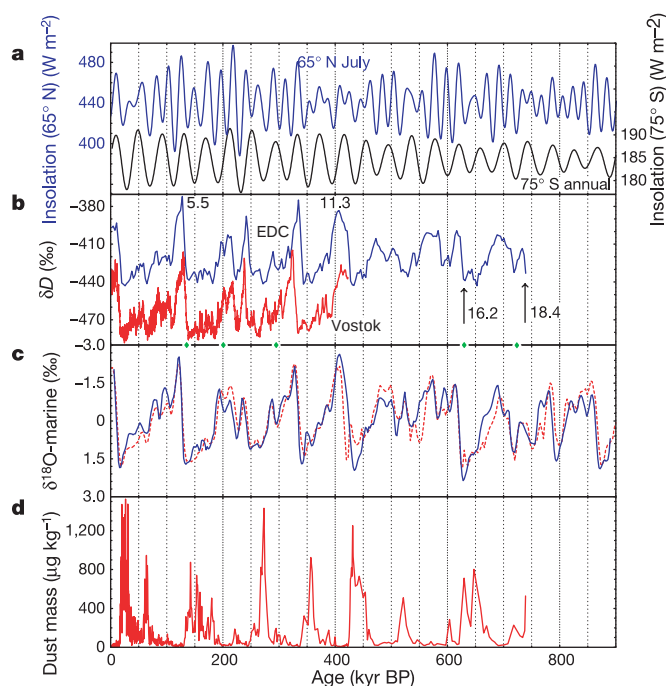


Figure 2 Comparison of EPICA Dome C data with other palaeoclimatic records. **a**, Insolation records⁴. Upper blue curve (left axis), mid-July insolation at 65° N; lower black curve (right axis), annual mean insolation at 75° S, the latitude of Dome C. **b**, δD from EPICA Dome C (3,000-yr averages). Vostok δD (red) is shown for comparison¹ and some MIS stage numbers are indicated; the locations of the control windows (below 800-m depth) used to make the timescale are shown as diamonds on the x-axis. **c**, Marine oxygen isotope record. The solid blue line is the tuned low-latitude stack of site MD900963 and ODP677³; to indicate the uncertainties in the marine records we also show (dashed red line) another record, which is a stack of seven sites for the last 400 kyr but consisting only of ODP site 677 for the earlier period². Both records have been normalized to their long-term average. **d**, Dust from EPICA Dome C.

obtained at Vostok, is that of Termination V. The integrity of this section can be tested using the depth difference expected between contemporaneous events recorded in the gas and the solid phase. We measured CO₂ and CH₄ mixing ratios in the air enclosed in the ice at ~1-m resolution between 2,760 and 2,800 m (Fig. 3). From the Vostok findings over the last four terminations²³, we expect the following pairs of events to be roughly synchronous: (1) the CO₂ peak/ δD peak, (2) the start of CO₂ increase/start of δD increase. The depth offset (Δ depth) values for these two pairs of 5 to 7 m are in reasonable agreement with Δ depth values calculated with the firn densification model, taking into account the thinning function obtained with the ice-flow model (Fig. 3). These observations support the conclusion that this part of the Dome C record is undisturbed, that is, that there is no folding of the ice.

Although visible ash layers tilted by a few degrees from the horizontal have been observed in the deeper ice, so far we have observed none of the highly inclined layers and overturned folds that were associated with stratigraphic disturbance in the lowest 10% of the deep Greenland (Summit) ice cores. The electrical records to 3,190 m also show no unexpectedly rapid changes, of the kind that might be diagnostic of folding. In conclusion, all the evidence supports the integrity of the ice-core stratigraphy to 3139 m.

Antarctic climate beyond MIS 11

One of the paradoxes of Quaternary climate is the dominance of 100-kyr periodicity in the past few climatic cycles, even though the amplitude of insolation changes at this period is rather small. This

can be addressed by examining changes in the amplitude and frequency of climate through the Quaternary period. On the basis mainly of ice-volume records, two major transitions have been identified. The mid-Pleistocene revolution (MPR) is characterized by an increase in mean global ice volume, and a change in the dominant period from 41 to 100 kyr (ref. 2). Its timing is often considered to be at about 900 kyr BP (that is, before the scope of this paper). A second distinct climate change, the mid-Brunhes event (MBE, for example²⁴), roughly corresponds to the transition between stage 12 and stage 11 (Termination V) about 430 kyr ago. The MBE is characterized by a further increase of ice-volume variations with, from then to the present day, four large-amplitude 100-kyr-dominated glacial–interglacial cycles. The intermediate period between the MPR and the MBE is characterized by a less-clear pattern. This schematic description of Quaternary climate, largely based on deep-sea isotopic records of ice-volume changes, also holds for at least some sea surface temperature records. For example, a composite South Atlantic 1,830-kyr record²⁵ shows cold and relatively stable summer temperatures before the MPR followed by higher-amplitude fluctuations between the MPR and the MBE and much stronger variations thereafter. Now we have the opportunity to examine the pre-MBE signal in Antarctic temperature and dust.

In the EDC δD record (Fig. 2), as in the marine-isotope records, the most striking feature is the greater amplitude of glacial–interglacial change in the period after Termination V (with 430 kyr as the boundary), compared to the earlier period. The standard deviation of the signals increases by 45% for EDC and 12% for the $\delta^{18}O$ record of ref. 3; other planktonic series show a similar feature²⁴. The Devil's Hole calcite isotopic record²⁶, which, however, extends only back to 565 kyr BP, also shows less variability before than after the MBE and indeed resembles the EDC record over the part common to both records. In detail, the period before Termination V in EDC is characterized by somewhat less cold glacial maxima (with the exception of stage 16.2), but by very significantly less warm interglacials (Fig. 4). Less extreme (weaker amplitude) interglacials occupied a larger proportion of each glacial/interglacial cycle, with the result that the mean δD value before and after 430 kyr is quite similar. The new ice-core data strongly emphasize the contrast in climate before and after the MBE.

The driving mechanisms for neither the MPR nor the MBE are as yet well understood. Some properties of the insolation curves have

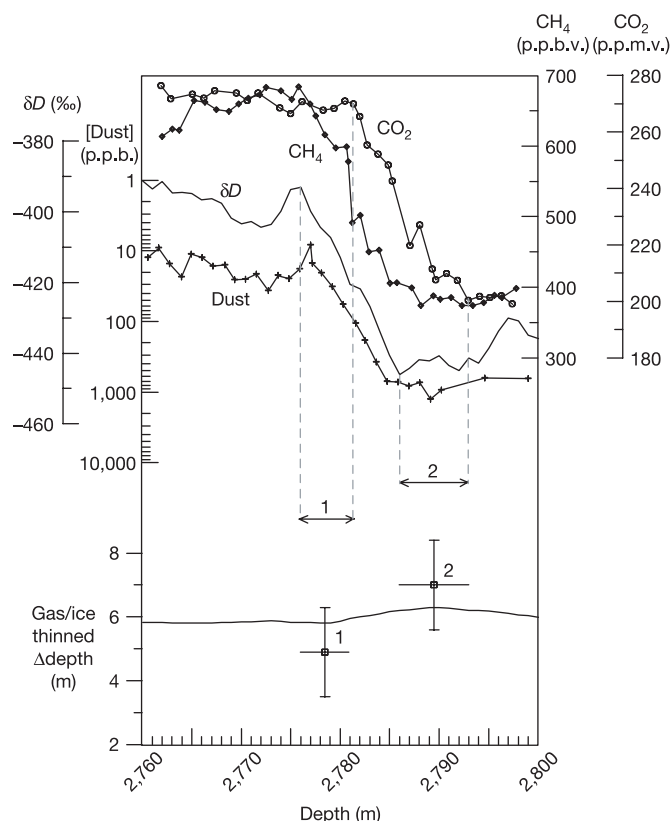


Figure 3 Termination V in the EPICA Dome C ice core on an ice depth scale. The top panel shows the ice-core parameters: circles, CO₂; diamonds, CH₄; line with no symbols, δD ; crosses, dust. The lower panel shows the modelled difference in depth between ice and air of the same age (line) along with estimates of the actual difference (error bars are based on uncertainty in aligning common events) for events considered roughly contemporaneous on the basis of their behaviour in later terminations at Vostok. Event 1, CO₂ peak/ δD peak; event 2, CO₂ early increase/ δD early increase.

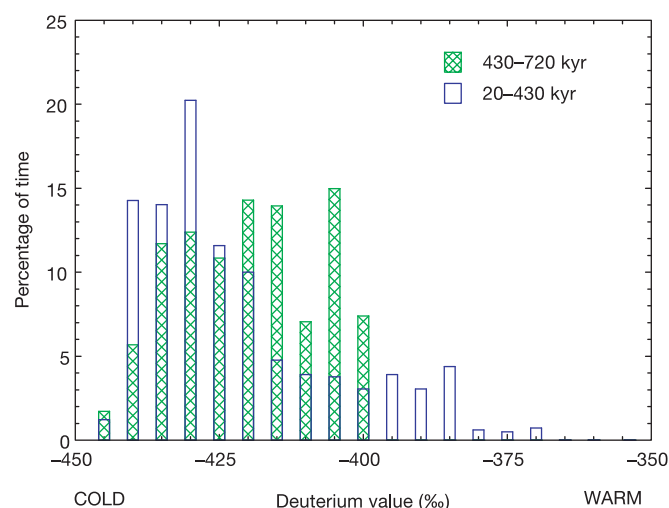


Figure 4 Histogram of δD values before and after 430 kyr. The bars show the occurrence of values within 5‰ windows for each of the periods, indicating that for the earlier period, there are no very warm values, but the time spent in warm and cold periods is more even than in the later period.

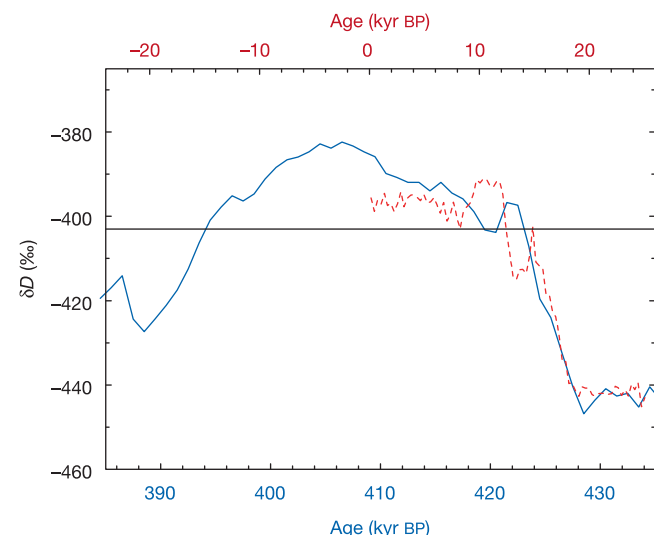


Figure 5 Comparison of Termination V plus MIS 11 with Termination I plus Holocene. δD data for MIS 11 (1-kyr averages) are shown as a solid blue line using the lower x-axis; data for the Holocene are shown as a dashed red line using the upper x-axis. Various alignments could be made, but we have adjusted the x-axes so that the start of each termination is aligned. A horizontal line is drawn at -403‰ .

changed progressively over the last 800 kyr, with an increased amplitude of obliquity changes, for example, and therefore an increased variability of annual local insolation (Fig. 2a) in the later part of the record. However, none of the simple conceptual models developed to simulate the timing of the Pleistocene glaciations has been able to suggest an explanation of the MBE. The climate became more orderly and predictable after the MBE, perhaps as a result of the emergence of new feedback mechanisms linked with changes in boundary conditions, such as the strength of ocean circulation, albedo, carbon dioxide or isostasy²⁴. At this stage, we have no additional clues allowing us to favour any one of these feedbacks, or to formulate other possibilities, but to obtain a detailed carbon dioxide record over 800,000 yr should certainly be helpful.

A final issue concerning the complete record is the stability in the size of the Antarctic ice sheet. Preliminary measurements of air content made between 2762.1 and 2783.0 m depth (MIS 11), between 3054.7 and 3059.1 m (MIS 16.3) and between 3099.8 and 3100.9 m (MIS 17.3) show the same mean value as that of EDC ice dating from the last 40 kyr ($0.089 \text{ cm}^3 \text{ g}^{-1}$). This suggests²⁷ that over the last 700 kyr, the surface elevation in this central part of East Antarctica has been as stable as during the last 40 kyr. This sets constraints, probably of the order of 5 m (ref. 28), on the possible contribution of this part of East Antarctica to changes in sea level²⁹.

Termination V

MIS 11 emerges as a key interglacial, both as viewed from the atmosphere in the EDC record and from the ocean in the $\delta^{18}\text{O}$

marine records. It delimits the frontier between two different patterns of climate, and has been identified as a unique and exceptionally long interglacial³⁰. Some authors suggest that, because the orbital parameters (low eccentricity and consequently weak precessional forcing) are similar to those of the present and the next tens of thousands of years, MIS 11 may be the best analogue for present and future climate without human intervention³¹. In this context, we note (Fig. 5) that, on the EDC2 timescale, δD (our temperature proxy) remains above -403‰ (the minimum 300-yr average value observed during the full Holocene epoch) for 28 kyr in MIS 11 (apart from a brief reversal near the start); in the Holocene, δD has so far been above -403‰ for 12 kyr. The rate of change in δD is very similar in Terminations V and I. Both terminations show a clear temperature reversal, but the one in the earlier period occurs after interglacial warmth has already been achieved. Thus the reversal at about 420 kyr might be seen as analogous to the Antarctic cold reversal (ACR) that occurred during Termination I at around 13 kyr, or it might be seen as similar to the dip (at about 8 kyr) that occurred after the early Holocene warm period.

Our low-resolution data for CO_2 , CH_4 (Fig. 3) and other parameters already provide information on how Termination V mimics or differs from younger terminations in terms of coupling between climate and greenhouse gases. With a minimum at 200 p.p.m.v. and 380 p.p.b.v. at the end of MIS 12 and a maximum at 275 p.p.m.v. and 680 p.p.b.v. at the start of MIS 11, CO_2 and CH_4 mixing ratios lie within the range observed during younger glacials and interglacials¹, the MIS 12 values being slightly at the higher end of the glacial range. These observations and the incomplete MIS 11 CO_2 record measured along the Vostok ice core²⁸ rule out unusual greenhouse conditions during MIS 11³² or a link between coral-reef growth and the intense carbonate dissolution of MIS 11 through unusual CO_2 mixing ratios³⁰. Other parameters measured on the core (Table 1), representing conditions and transport in different compartments of the environment, have very similar (glacial) values at equivalent points just before Terminations I and V and very similar (interglacial) values just after the two transitions. This confirms that, in all the proxies we are able to examine, there is no significant long-term trend in the period since the MBE.

The general shape of the greenhouse gas increases resembles younger terminations, that is, a regular trend for CO_2 and a two-step transition for CH_4 (slow increase followed by a rapid jump towards interglacial values); however, no Younger-Dryas-like event is observed in our CH_4 profile.

The most striking feature concerns the relative timing of the CO_2 and CH_4 increases compared with younger terminations: whereas CH_4 started to increase concomitantly with CO_2 (and Antarctic temperature) during the last four terminations, at Termination V it leaves its glacial background 4 to 5 kyr later than CO_2 , by which time the latter had already increased by about 50 p.p.m.v. Similarly, the rapid jump of CH_4 punctuating the second part of its transition takes place when CO_2 approaches its maximum. Note that this is also the time when Antarctic temperature starts a slow decrease, that is, a typical expression of a bipolar see-saw as observed during stage 3 (ref. 33) and possibly Termination I¹². Following its rapid jump at

Table 1 Concentrations of major analytes measured along the EDC ice core

Analytes	6–8 kyr BP (after Termination I)	20–22 kyr BP (before Termination I)	416–418 kyr BP (after Termination V)	430–432 kyr BP (before Termination V)
δD (‰)	–399	–442	–395	–442
CO_2 (p.p.m.v.)	260	185	270	200
CH_4 (p.p.b.v.)	600	360	670	390
Dust ($\mu\text{g kg}^{-1}$)	14	680	24	630
Na ($\mu\text{g kg}^{-1}$)	20	101	23	99
SO_4^{2-} ($\mu\text{g kg}^{-1}$)	92	201	98	216

Gas values are for Dome C. No corrections for interhemispheric differences or global averages have been applied. Data is shown for approximately equivalent periods before and after Terminations I and V.

the end of the termination, CH₄ continues to increase by ~100 p.p.b.v. for 2 to 3 kyr, another unusual feature when compared to the CH₄ trends during the early part of MIS 1, 5, 7 and 9 (ref. 1).

A thorough discussion of the causes of these greenhouse-gas peculiarities during Termination V is beyond the scope of this paper. But evidently the similarities and differences observed with younger terminations will stimulate the debate on how greenhouse gas and climate are coupled on Quaternary timescales.

Prospects from the rest of the core

In this paper, we have shown the extended climate record back to 740 kyr, and that the pattern of climate before MIS 11 was different to that which has followed for the past four glacial cycles. Although the results from MIS 11 indicate that without human intervention a climate similar to the present one would extend well into the future, the predicted increases in greenhouse-gas concentrations make this unlikely³⁴.

According to our preliminary timescale, extending the record to 3,190 m (ice already drilled but not analysed) will take the record back to 807 ± 10 kyr (MIS 20.2). The electrical records already obtained on this ice (Fig. 1), although difficult to interpret simply in terms of climate, certainly suggest that another glacial cycle will be found in this ice. This ice should include the Brunhes–Matuyama magnetic reversal, generally dated to about 780 kyr, and therefore give us the first indication of how a reversal is recorded in cosmogenic isotopes such as ¹⁰Be.

There remains up to 120 m of ice still to drill. This will be difficult to obtain because the ice is near to the melting temperature. The timescale EDC2 extended to the base gives an age of 960 ± 20 kyr. Therefore, when the record is complete, we could expect to reach MIS 26 (just beyond the MPR), assuming that the integrity of the stratigraphy and all the approximations of the dating method are still reasonable down to the base. It will be of particular interest to see how the tight coupling between greenhouse gases and Antarctic temperature (δD) seen in the last 420 kyr evolves through the earlier parts of the record. □

Methods

Analysis

The electrical conductivity measurement determines the d.c. conductance between electrodes on a fresh ice surface. Dielectric profiling determines the conductivity of the ice at higher frequencies. Both were measured in the field at a temperature of −20 ± 2 °C, corrected^{15,16} to −15 °C. Data were collected at high resolution and averaged to 1 m. Vertical thin sections were prepared in the field at a periodicity of 10 m, then digitized and analysed using an image analysis procedure³⁵ to determine the mean grain radius. A 3.4 cm × 3.4 cm strip of ice was melted on a hotplate in the field³⁶, and fed into various detectors. Aliquots (1.1-m averages) were also collected from this melting device into clean containers, frozen and shipped to Europe for ion chromatographic analysis³⁷ of major ions (presented for Termination V). All other measurements were made in laboratories in Europe after the ice had been shipped frozen from Dome C. δD was determined¹⁰ on meltwater from 55-cm-long sections. This record, still discontinuous for some parts, should be considered as preliminary. Also, we used a ‘quick’ mode (each sample is measured twice instead of four times), leading to a typical accuracy of 1.5‰ (1σ), whereas we aim for a final precision of 0.5‰ over the entire core, as currently obtained for EDC96 (the upper 780 m). δD data shown in Fig. 1 correspond to values averaged across seven successive samples. The current precision and resolution are well adapted for the climatic interpretation discussed here (Fig. 2), in which we focus on the broad features of Antarctic climate changes over the past eight climatic cycles.

Dust concentration and size distribution was measured by a 256-channel Coulter Counter, set to register particles in the size range from 0.7–20 μm (ref. 13). In calculating mass concentrations, density was taken as 2,500 kg m^{−3}. CO₂ and CH₄ were measured (for Termination V) by a dry crushing¹² and a melt-refreezing extraction technique³⁸, respectively.

Models used for ice-core dating

Full details of the derivation of the timescale are given in the Supplementary Information. For the thinning rate computation, we used an ice-flow model³⁹, with prescribed surface elevation⁴⁰. It has two poorly known parameters: the melting at the base of the ice sheet (*F*), which is the condition for the vertical velocity at the base, and a parameter (*m*) for the vertical velocity profile. The vertical strain rate is assumed to be proportional to 1 − (*z/H*)^(*m*+1), where *z* is the depth and *H* is the ice thickness. The accumulation rate is deduced from the δD content of the ice, via the temperature of the inversion layer. This

conversion involves two further tunable parameters. The last modelling step of the chronology is the evaluation of the difference between the gas age and the ice age (Δage), which is required to derive the age scale for the gas measurements. This is derived from a firm model²². The four poorly known parameters of the models are evaluated through the use of a small number of chronological controls, through a Monte Carlo inverse method^{3,21}. The method searches for an optimal agreement, within the limits of the confidence interval of each assigned age (that is, we use control windows rather than control points) and using the same rules to define accumulation all along the record. In the top part of the core, we use the same control points as were used to derive the timescale (EDC1) recommended for the shallower part of the core⁴¹; EDC1 remains the recommended timescale for this part of the core, and hands over precisely to EDC2 at 800 m. For the bottom part of the core (that is, for the period older than 50 kyr), we used several age control windows derived by comparison to the stacked marine isotope curve of Bassinot³, assuming a 4-kyr phase lag. These points are situated at Terminations II (1,738 m = 131 ± 6 kyr), III (2,311 m = 245 ± 6 kyr), IV (2,593 m = 338 ± 6 kyr), VII (3,038 m = 626 ± 6 kyr) and VIII (3,119 m = 717 ± 6 kyr). Note that the age of identical events in this EDC2 chronology can differ, over their common parts, from the Vostok and Dome Fuji chronologies, because of slightly different best-fit parameters in the model.

Received 27 February; accepted 22 April 2004; doi:10.1038/nature02599.

- Petit, J. R. *et al.* Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–436 (1999).
- Imbrie, J. *et al.* On the structure and origin of major glaciation cycles. 2. The 100,000-year cycle. *Paleoceanography* **8**, 699–735 (1993).
- Bassinot, F. C. *et al.* The astronomical theory of climate and the age of the Brunhes–Matuyama magnetic reversal. *Earth Planet. Sci. Lett.* **126**, 91–108 (1994).
- Berger, A. & Loutre, M. F. Insolation values for the climate of the last 10 million years. *Quat. Sci. Rev.* **10**, 297–317 (1991).
- Watanabe, O. *et al.* Homogeneous climate variability across East Antarctica over the past three glacial cycles. *Nature* **422**, 509–512 (2003).
- Fischer, H., Wahlen, M., Smith, J., Mastoianni, D. & Deck, B. Ice core records of atmospheric CO₂ around the last three glacial terminations. *Science* **283**, 1712–1714 (1999).
- Johnsen, S. J. *et al.* Oxygen isotope and palaeotemperature records from six Greenland ice-core stations: Camp Century, Dye-3, GRIP, GISP2, Renland and NorthGRIP. *J. Quat. Sci.* **16**, 299–307 (2001).
- The EPICA Dome C 2001–02 science and drilling teams, Extending the ice core record beyond half a million years. *Eos Trans.* **83**, 509–517 (2002).
- Lorius, C., Merlivat, L., Jouzel, J. & Pourchet, M. A 30,000-yr isotope climatic record from Antarctic ice. *Nature* **280**, 644–648 (1979).
- Jouzel, J. *et al.* A new 27 ky high resolution East Antarctic climate record. *Geophys. Res. Lett.* **28**, 3199–3202 (2001).
- Stenni, B. *et al.* An oceanic cold reversal during the last deglaciation. *Science* **293**, 2074–2077 (2001).
- Monnin, E. *et al.* Atmospheric CO₂ concentrations over the last glacial termination. *Science* **291**, 112–114 (2001).
- Delmonte, B., Petit, J. R. & Maggi, V. Glacial to Holocene implications of the new 27000-year dust record from the EPICA Dome C (East Antarctica) ice core. *Clim. Dyn.* **18**, 647–660 (2002).
- Röthlisberger, R. *et al.* Dust and sea salt variability in central East Antarctica (Dome C) over the last 45 kyrs and its implications for southern high-latitude climate. *Geophys. Res. Lett.* **29**, doi:10.1029/2002GL015186 (2002).
- Stauffer, B., Flückiger, J., Wolff, E. W. & Barnes, P. R. F. The EPICA deep ice cores: First results and perspectives. *Ann. Glaciol.* (in the press).
- Wolff, E. W., Basile, I., Petit, J.-R. & Schwander, J. Comparison of Holocene electrical records from Dome C and Vostok, Antarctica. *Ann. Glaciol.* **29**, 89–93 (1999).
- Weiss, J. *et al.* Dome Concordia ice microstructure: impurities effect on grain growth. *Ann. Glaciol.* **35**, 552–558 (2002).
- Jouzel, J. *et al.* Magnitude of isotope/temperature scaling for interpretation of central Antarctic ice cores. *J. Geophys. Res.* **108**, doi:10.1029/2002JD002677 (2003).
- Blunier, T., Schwander, J., Chappellaz, J., Parrenin, F. & Barnola, J. M. What was the surface temperature in central Antarctica during the last glacial maximum? *Earth Planet. Sci. Lett.* **218**, 379–388 (2004).
- Röthlisberger, R. *et al.* Limited dechlorination of sea salt aerosols during the last glacial period—Evidence from the European Project for Ice Coring in Antarctica (EPICA) Dome C ice core. *J. Geophys. Res.* **108**, doi:10.1029/2003JD003604 (2003).
- Parrenin, F., Jouzel, J., Waelbroeck, C., Ritz, C. & Barnola, J. M. Dating the Vostok ice core by an inverse method. *J. Geophys. Res.* **106**, 31837–31851 (2001).
- Goujon, C., Barnola, J. M. & Ritz, C. Modeling the densification of polar firn including heat diffusion: Application to close-off characteristics and gas isotopic fractionation for Antarctica and Greenland sites. *J. Geophys. Res.* **108**, doi:10.1029/2002JD003319 (2003).
- Pepin, L., Raynaud, D., Barnola, J. M. & Loutre, M. F. Hemispheric roles of climate forcings during glacial–interglacial transitions as deduced from the Vostok record and LLN-2D model experiments. *J. Geophys. Res.* **106**, 31885–31892 (2001).
- Berger, W. H. & Wefer, G. in *Earth's Climate and Orbital Eccentricity: the Marine Isotope Stage 11 Question* Geophys. Monogr. 137 (eds Droxler, A. W., Poore, R. Z. & Burckle, L. H.) 41–59 (AGU, Washington, 2003).
- Beccquey, S. & Gersonde, R. Past hydrographic and climatic changes in the Subantarctic Zone of the South Atlantic—The Pleistocene record from ODP Site 1090. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **182**, 221–239 (2002).
- Winograd, I. J. *et al.* Continuous 500,000-year climate record from vein calcite in Devils-Hole, Nevada. *Science* **258**, 255–260 (1992).
- Martinerie, P., Raynaud, D., Etheridge, D. M., Barnola, J.-M. & Mazaudier, D. Physical and climatic parameters which influence the air content in polar ice. *Earth Planet. Sci. Lett.* **112**, 1–13 (1992).
- Raynaud, D., *et al.* in *Earth's Climate and Orbital Eccentricity: the Marine Isotope Stage 11 Question* Geophys. Monogr. 137 (eds Droxler, A. W., Poore, R. Z. & Burckle, L. H.) 27–40 (AGU, Washington, 2003).
- Hearty, P. J., Kindler, P., Cheng, H. & Edwards, R. L. A +20 m middle Pleistocene sea-level highstand

- (Bermuda and the Bahamas) due to partial collapse of Antarctic ice. *Geology* **27**, 375–378 (1999).
30. Droxler, A. W., Alley, R. B., Howard, W. R., Poore, R. Z. & Burckle, L. H. in *Earth's Climate and Orbital Eccentricity: the Marine Isotope Stage 11 Question* Geophys. Monogr. 137 (eds Droxler, A. W., Poore, R. Z. & Burckle, L. H.) 1–14 (AGU, Washington, 2003).
 31. Berger, A. & Loutre, M. F. in *Earth's Climate and Orbital Eccentricity: the Marine Isotope Stage 11 Question* Geophys. Monogr. 137 (eds Droxler, A. W., Poore, R. Z. & Burckle, L. H.) 17–26 (AGU, Washington, 2003).
 32. Lea, D. W., Pak, D. K. & Spero, H. J. in *Earth's Climate and Orbital Eccentricity: the Marine Isotope Stage 11 Question* Geophys. Monogr. 137 (eds Droxler, A. W., Poore, R. Z. & Burckle, L. H.) 147–156 (AGU, Washington, 2003).
 33. Stocker, T. F. & Johnsen, S. J. A minimum thermodynamic model for the bipolar seesaw. *Paleoceanography* **18**, doi:10.1029/2003PA000920 (2003).
 34. IPCC, *IPCC Third Assessment Report: Climate Change 2001: The Scientific Basis* (Cambridge Univ. Press, Cambridge, 2001).
 35. Gay, M. & Weiss, J. Automatic reconstruction of polycrystalline ice microstructure from image analysis: application to the EPICA ice core at Dome Concordia, Antarctica. *J. Glaciol.* **45**, 547–554 (1999).
 36. Röthlisberger, R. *et al.* Technique for continuous high-resolution analysis of trace substances in firm and ice cores. *Environ. Sci. Technol.* **34**, 338–342 (2000).
 37. Littot, G. C. *et al.* Comparison of analytical methods used for measuring major ions in the EPICA Dome C (Antarctica) ice core. *Ann. Glaciol.* **35**, 299–305 (2002).
 38. Landais, A. *et al.* A tentative reconstruction of the last interglacial and glacial inception in Greenland based on new gas measurements in the Greenland Ice Core Project (GRIP) ice core. *J. Geophys. Res.*

108, doi:10.1029/2002JD003147 (2003).

39. Ritz, C. *Un Modèle Thermo-Mécanique d'Évolution pour le Bassin Glaciaire Antarctique Vostok-Glacier Byrd: Sensibilité aux Valeurs des Paramètres Mal Connus* Thesis, Univ. J. Fourier (1992).
40. Ritz, C., Rommelaere, V. & Dumas, C. Modeling the evolution of Antarctic ice sheet over the last 420,000 years: Implications for altitude changes in the Vostok region. *J. Geophys. Res.* **106**, 31943–31964 (2001).
41. Schwander, J. *et al.* A tentative chronology for the EPICA Dome Concordia ice core. *Geophys. Res. Lett.* **28**, 4243–4246 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature

Acknowledgements We thank the logistics and drilling teams. This work is a contribution to the European Project for Ice Coring in Antarctica (EPICA), a joint European Science Foundation/European Commission (EC) scientific programme, funded by the EC and by national contributions from Belgium, Denmark, France, Germany, Italy, The Netherlands, Norway, Sweden, Switzerland and the UK.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.W. (ewwo@bas.ac.uk).

EPICA community members* (participants are listed alphabetically)

Laurent Augustin¹, Carlo Barbante², Piers R. F. Barnes³, Jean Marc Barnola¹, Matthias Bigler⁴, Emiliano Castellano⁵, Olivier Cattani⁶, Jerome Chappellaz¹, Dorte Dahl-Jensen⁷, Barbara Delmonte^{1,8}, Gabrielle Dreyfus⁶, Gael Durand¹, Sonia Falourd⁶, Hubertus Fischer⁹, Jacqueline Flückiger⁴, Margareta E. Hansson¹⁰, Philippe Huybrechts⁹, Gérard Jugie¹¹, Sigfus J. Johnsen⁷, Jean Jouzel⁶, Patrik Kaufmann⁴, Josef Kipfstuhl⁹, Fabrice Lambert⁴, Vladimir Y. Lipenkov¹², Geneviève C. Littot³, Antonio Longinelli¹³, Reginald Lorrain¹⁴, Valter Maggi⁸, Valerie Masson-Delmotte⁶, Heinz Miller⁹, Robert Mulvaney³, Johannes Oerlemans¹⁵, Hans Oerter⁹, Giuseppe Orombelli⁸, Frederic Parrenin^{1,6}, David A. Peel³, Jean-Robert Petit¹, Dominique Raynaud¹, Catherine Ritz¹, Urs Ruth⁹, Jakob Schwander⁴, Urs Siegenthaler⁴, Roland Souchez¹⁴, Bernhard Stauffer⁴, Jorgen Peder Steffensen⁷, Barbara Stenni¹⁶, Thomas F. Stocker⁴, Ignazio E. Tabacco¹⁷, Roberto Udisti⁵, Roderik S. W. van de Wal¹⁵, Michiel van den Broeke¹⁵, Jerome Weiss¹, Frank Wilhelms⁹, Jan-Gunnar Winther¹⁸, Eric W. Wolff³ & Mario Zucchelli^{19*}

1, Laboratoire de Glaciologie et Géophysique de l'Environnement (CNRS), BP 96, 38402 St Martin d'Hères Cedex, France; 2, Environmental Sciences Department, University of Venice, Calle Larga S. Marta, 2137, I-30123 Venice, Italy; 3, British Antarctic Survey, High Cross, Madingley Road, Cambridge CB3 0ET, UK; 4, Climate and Environmental Physics, Physics Institute, University of Bern, Sidlerstrasse 5, CH-3012 Bern, Switzerland; 5, Department of Chemistry—Analytical Chemistry Section, Scientific Pole—University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino (Florence), Italy; 6, Institut Pierre Simon Laplace/Laboratoire des Sciences du Climat et de l'Environnement, UMR CEA-CNRS 1572, CE Saclay, Orme des Merisiers, 91191 Gif-Sur-Yvette, France; 7, Niels Bohr Institute for Astronomy, Physics and Geophysics, University of Copenhagen, Juliane Maries Vej 30, DK-2100 Copenhagen, Denmark; 8, University of Milano-Bicocca, Dipartimento di Scienze Ambiente e Territorio, Piazza della Scienza 1, I-20126 Milan, Italy; 9, Alfred-Wegener-Institute for Polar- und Marine Research (AWI), Postfach 120161, D-27515 Bremerhaven, Germany; 10, Department of Physical Geography and Quaternary Geology, Stockholm University, S-106 91 Stockholm, Sweden; 11, Institut Polaire Français—Paul Emile Victor (IPEV), BP 75, 29280 Plouzané, France; 12, Arctic and Antarctic Research Institute, 38 Beringa Street, 199397 St Petersburg, Russia; 13, Department of Earth Sciences, University of Parma, Parco Area delle Scienze 157/A, I-43100 Parma, Italy; 14, Département des Sciences de la Terre et de l'Environnement, Faculté des Sciences, CP 160/03, Université Libre de Bruxelles, 50 avenue FD Roosevelt, B1050 Brussels, Belgium; 15, Institute for Marine and Atmospheric Research Utrecht (IMAU), Princetonplein 5, 3584 CC Utrecht, The Netherlands; 16, Department of Geological, Environmental and Marine Sciences, University of Trieste, Via E. Weiss 2, I-34127 Trieste, Italy; 17, Earth Science Department, University of Milan, Via Cicognara 7, 20129 Milano, Italy; 18, Norwegian Polar Institute, N-9296 Tromsø, Norway; 19, ENEA, CRE Casaccia, PO Box 2400, Via Anguillarese 301, 00060 S. Maria di Galeria (RM), Italy.

*Deceased.

Subtractive proteomic mapping of the endothelial surface in lung and solid tumours for tissue-specific therapy

Phil Oh, Yan Li, Jingyi Yu, Eberhard Durr, Karolina M. Krasinska, Lucy A. Carver, Jacqueline E. Testa & Jan E. Schnitzer

Sidney Kimmel Cancer Center, 10835 Altman Row, San Diego, California 92121, USA

The molecular complexity of tissues and the inaccessibility of most cells within a tissue limit the discovery of key targets for tissue-specific delivery of therapeutic and imaging agents *in vivo*. Here, we describe a hypothesis-driven, systems biology approach to identifying a small subset of proteins induced at the tissue–blood interface that are inherently accessible to antibodies injected intravenously. We use subcellular fractionation, subtractive proteomics and bioinformatics to identify endothelial cell surface proteins exhibiting restricted tissue distribution and apparent tissue modulation. Expression profiling and γ -scintigraphic imaging with antibodies establishes two of these proteins, aminopeptidase-P and annexin A1, as selective *in vivo* targets for antibodies in lungs and solid tumours, respectively. Radio-immunotherapy to annexin A1 destroys tumours and increases animal survival. This analytical strategy can map tissue- and disease-specific expression of endothelial cell surface proteins to uncover novel accessible targets useful for imaging and therapy.

New targets are needed for detecting disease through molecular imaging^{1–4} and for treating disease through directed delivery *in vivo*^{5–9}. Sequencing the human genome has identified a target pool of 40,000 genes which may translate into a million possible protein targets¹⁰. Genomic and proteomic analysis of normal and diseased tissues has yielded thousands of genes and gene products for diagnostic and tissue assignment as well as potential therapeutic targets^{5–11}. But the sheer number of candidates can overwhelm the required *in vivo* validation process, leading some to question the ultimate impact of these approaches on speeding up drug discovery^{5–7,10}. Reducing tissue data complexity to a manageable subset of candidates most relevant to targeting, imaging and treating disease is clearly desired but requires new discovery and validation strategies, that effectively focus the power of global identification technologies.

Selectively targeting a single organ or diseased tissue (such as a solid tumour) *in vivo* remains a desirable but elusive goal for molecular medicine^{12,13}. This targeting would allow more effective imaging as well as drug and gene therapies for many acquired and genetic diseases^{1–4,6,7,10–13}. Most tissue- and disease-associated proteins are expressed by cells inside tissue compartments not readily accessible to intravenously injected biological agents such as antibodies. This inaccessibility hinders many site-directed therapies^{5–7,12,13} and imaging agents^{1–4}. For example, multiple barriers to delivery prevent effective immunotherapy for solid tumours *in vivo*, despite efficacy and specificity *in vitro*^{12–17}. Conversely, the nearly universal access of chemotherapeutics dilutes efficacy requiring increased dosages, leading to unwanted systemic side effects. Thus, new approaches are required to filter the overabundance of molecular information to allow rapid discovery and validation of accessible tissue-specific targets that can direct molecular imaging and pharmacodelivery *in vivo*.

Here, we divert analytical focus from the poorly accessible cells inside tissues to the critical tissue–blood interface, inherently accessible to agents absorbed or injected into the blood. The endothelium is a minor tissue component that exists as an attenuated cell monolayer lining all blood vessels and is in direct contact with the circulating blood. The tissue microenvironment surrounding blood vessels seems to control the endothelial cell phenotype

in vivo^{18–22}. But the degree to which normal and diseased tissues can modulate endothelial cell surface protein expression is unknown. Indirect evidence of endothelial cell molecular heterogeneity comes from the reported ability of certain cells and peptide sequences to home to specific tissues after intravenous injection^{23–25}. Several receptors to angiogenic factors have been identified^{26–28} and genomic analysis of endothelial cells isolated from tumours has provided angiogenic markers of unknown subcellular localization²⁹. Genetically induced endothelial cell targets provide ‘proof of principle’ that immuno-targeting endothelium can cause tumour regression³⁰; however, few probes to native endothelial cell surface proteins show validated tissue-specific pharmacodelivery *in vivo*^{14,15,17,24,31,32}. This paucity of molecular information may reflect the lack of tissue-induced proteins or, perhaps more likely, the difficulty in studying such a minor cell type as it exists natively in the tissue without tissue disruption and isolation. Although gene and immune targeting of tumour neovasculature seems promising^{33,34}, the ultimate potential of endothelial targeting in creating new drug delivery systems remains unclear^{16,17}. The actual existence of tissue-specific endothelial cell surface targets will remain uncertain until we overcome current technical barriers to comprehensively map and validate tissue-induced endothelial cell surface proteins *in vivo*.

Molecular diversity of endothelium *in vivo*

To investigate endothelial cell surface heterogeneity *in vivo*, we performed subcellular fractionation to isolate silica-coated luminal endothelial cell plasma membranes and their caveolae directly from normal organs^{35,36}. These plasma membranes and their caveolae displayed a greater than 20-fold enrichment for endothelial cell surface or caveolar markers (angiotensin converting enzyme (ACE), vascular endothelial (VE)-cadherin and caveolin-1), whereas markers of intracellular organelles (for example, the Golgi protein β -COP), other tissue cells (for example, E-cadherin for epithelium, fibroblast surface protein) and blood (for example, glycophorin A, CD4 and CD11) were greater than 20-fold depleted (Fig. 1c; data not shown and as shown previously^{19,35,36}). This quality control analysis was applied to each isolate.

We analysed these isolated plasma membranes by two-

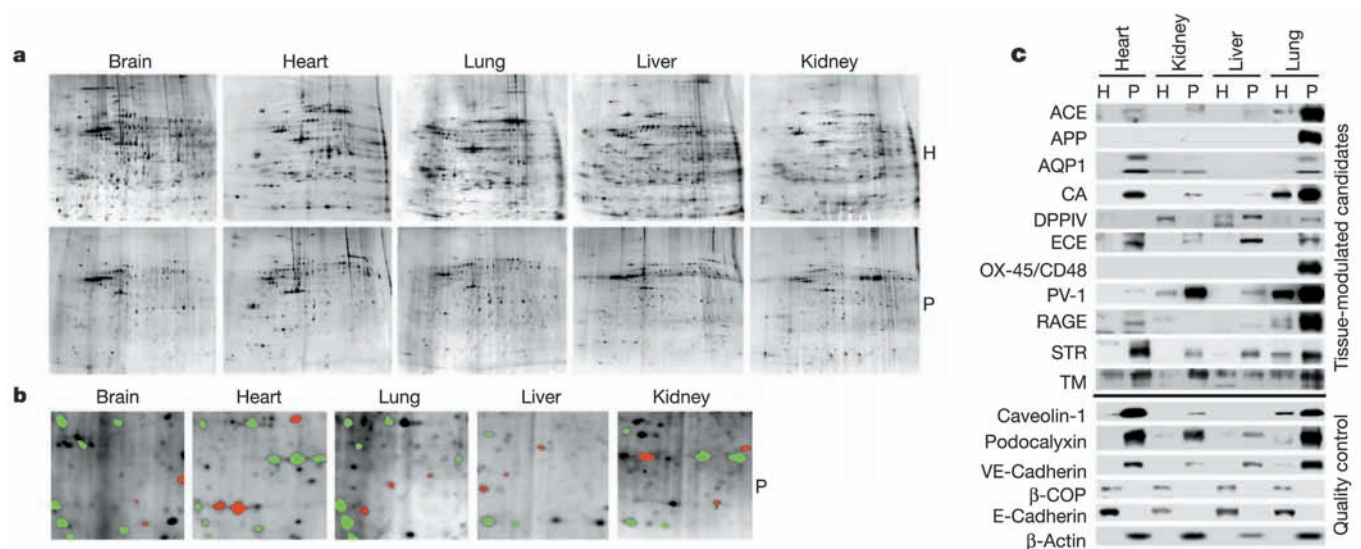


Figure 1 Molecular heterogeneity of endothelial cell surface in multiple rat organs *in vivo*. **a**, Silver-stained 2D gels of proteins from the whole tissue homogenates (H) and the isolated luminal cell endothelial plasma membranes (P). **b**, Comparison of representative region of 2D gels of endothelial plasma membrane from the indicated rat tissues. Green, protein spots found in ≥ 2 tissues; Red, spots not detected elsewhere. **c**, Western blot analysis of proteins from endothelial plasma membranes (10 μ g) from multiple rat organs

using the antibodies to the indicated proteins. ACE, angiotensin converting enzyme; APP, aminopeptidase P; AQP1, aquaporin 1; CA, carbonic anhydrase; DPPIV, dipeptidyl peptidase IV; ECE, endothelin converting enzyme; RAGE, receptor for advanced glycation end products; STR, seven transmembrane receptor; TM, thrombomodulin. β -COP, coatomer beta subunit.

dimensional (2D) gel electrophoresis to produce high-resolution vascular endothelial protein maps of the major rat organs, that were distinct and much reduced in complexity from that of the starting tissue homogenate (Fig. 1a). Differential spot analysis detected many distinct proteins in plasma membrane versus the homogenate and also in the plasma membrane between organs (Fig. 1b). Western blot analysis further confirmed this heterogeneity (Fig. 1c; data not shown). In many cases, antigens difficult to detect in tissue homogenates were readily apparent in the isolated plasma membranes, reflecting the significant enrichment and increased sensitivity provided by subfractionation to unmask proteins located on the endothelial cell surface. Unique 'molecular fingerprints or signatures' that may include tissue- and cell-specific proteins were apparent for each endothelium.

Subtractive analysis and profiling

To identify specific proteins expressed at the endothelial cell surface we used mass spectrometry, database searching and immunoblotting, to analyse endothelial plasma membranes and caveolae isolated from rat tissues^{35,36}, as well as cultured rat lung microvascular endothelial cells (RLMVEC), isolated and grown in culture³⁷. So far, we have identified nearly 2,000 non-redundant proteins (our unpublished observation) which were entered into a database (Accessible Vascular Targets, AVATAR) that was designed and annotated to provide relational analysis between measurements and samples. To identify lung-induced, and possibly, lung-specific endothelial cell surface proteins we used a systems biology approach, based on the hypothesis that the surrounding tissue microenvironment (normal or diseased) modulates protein expression in the vascular endothelium. We subtracted, *in silico*, the proteins identified in endothelial plasma membranes from rat lungs versus RLMVEC because tissue-dependent protein expression may be reduced in cell culture that cannot yet duplicate conditions *in vivo*. To increase certainty of expression *in vivo* versus *in vitro*, we applied a strict identification criterion of ≥ 6 tandem mass spectrometry (MS/MS) spectra to rat lung endothelial plasma membranes versus low stringency of ≥ 1 MS/MS spectra to RLMVEC

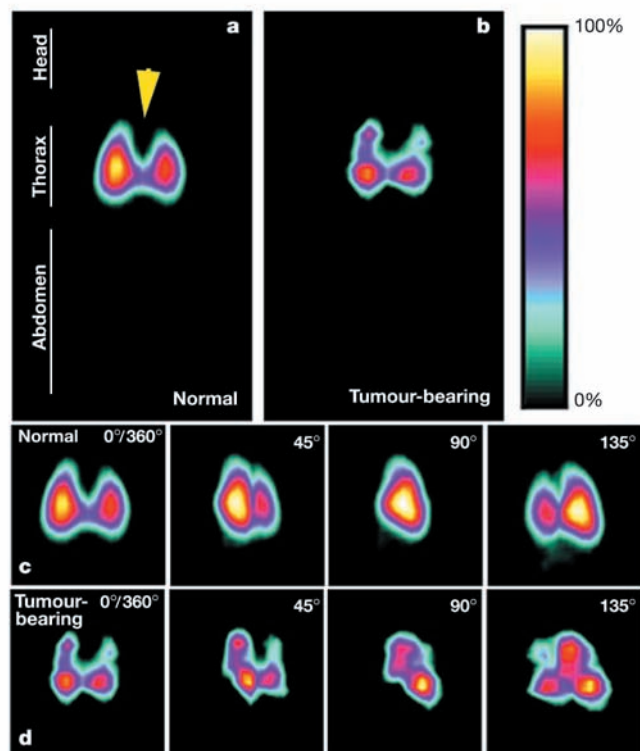


Figure 2 *In vivo* γ -scintigraphic imaging of APP antibody targeting in normal and tumour-bearing rats. **a**, **b**, Planar images collected 20 min after an intravenous injection of ^{125}I -APP monoclonal antibody into normal female rats (**a**) and female rats bearing metastatic breast adenocarcinoma in the lung (**b**). **c**, **d**, SPECT imaging using a parallel hole collimator of female normal rats (**c**) or rats bearing breast adenocarcinoma tumours in their lungs (**d**) 20 min after tail vein injection with ^{125}I -APP antibodies (50 μCi of IgG at 10 Ci g^{-1}). Representative re-projected frames at the indicated degree of rotation are shown. Arrowhead in (**a**) denotes heart cavity.

plasma membranes. From the current rat database of approximately 40,000 entries, we identified 37 differentially expressed proteins. To meet our objective of direct exposure to circulating antibodies, we applied bioinformatic interrogation of structure, glycosylation and membrane orientation to reduce the list to 11 proteins likely to have domains outside the cell.

To confirm the original mass spectrometry findings and to assess organ-specificity, we performed expression profiling with specific antibodies. Consistent with our hypothesis, western blot analysis detected all 11 proteins enriched in lung endothelial plasma membranes (Fig. 1c) but not in RLMVEC plasma membranes (data not shown). Each of these proteins exhibited a different pattern of restricted expression. Two proteins, aminopeptidase P (APP) and OX-45, were detected only in endothelial plasma membranes from lung. Staining of tissue sections with antibodies (including to APP and OX-45) confirmed this restricted expression (data not shown). This profiling by western blot analysis was very sensitive and reliable, even at low expression levels. Thus subtractive

proteomic analysis provided a subset of endothelial cell proteins not expressed *in vitro* and differentially expressed *in vivo*, apparently regulated by the unique tissue microenvironment in each organ.

Lung-specific targeting *in vivo*

To assess the immuno-accessibility of APP and OX-45 *in vivo*, including possible lung-specific targeting, we intravenously injected ^{125}I -labelled monoclonal antibodies into rats and performed whole-body imaging using planar γ -scintigraphy. Within 20 min, we observed clear lung images, indicating rapid and specific targeting of APP antibody to the lung (Fig. 2a, c). Region-of-interest and biodistribution analysis confirmed this lung-specific targeting with 65% of the injected dose per gram of tissue accumulating in the lungs and less than 2% injected dose per gram of tissue in other organs and blood. Consistent with the low blood levels, the heart cavity was readily apparent (Fig. 2a, arrowhead). Furthermore, this lung targeting was inhibited by addition of a 50-fold excess of unlabelled APP IgG but not with control IgG (data not

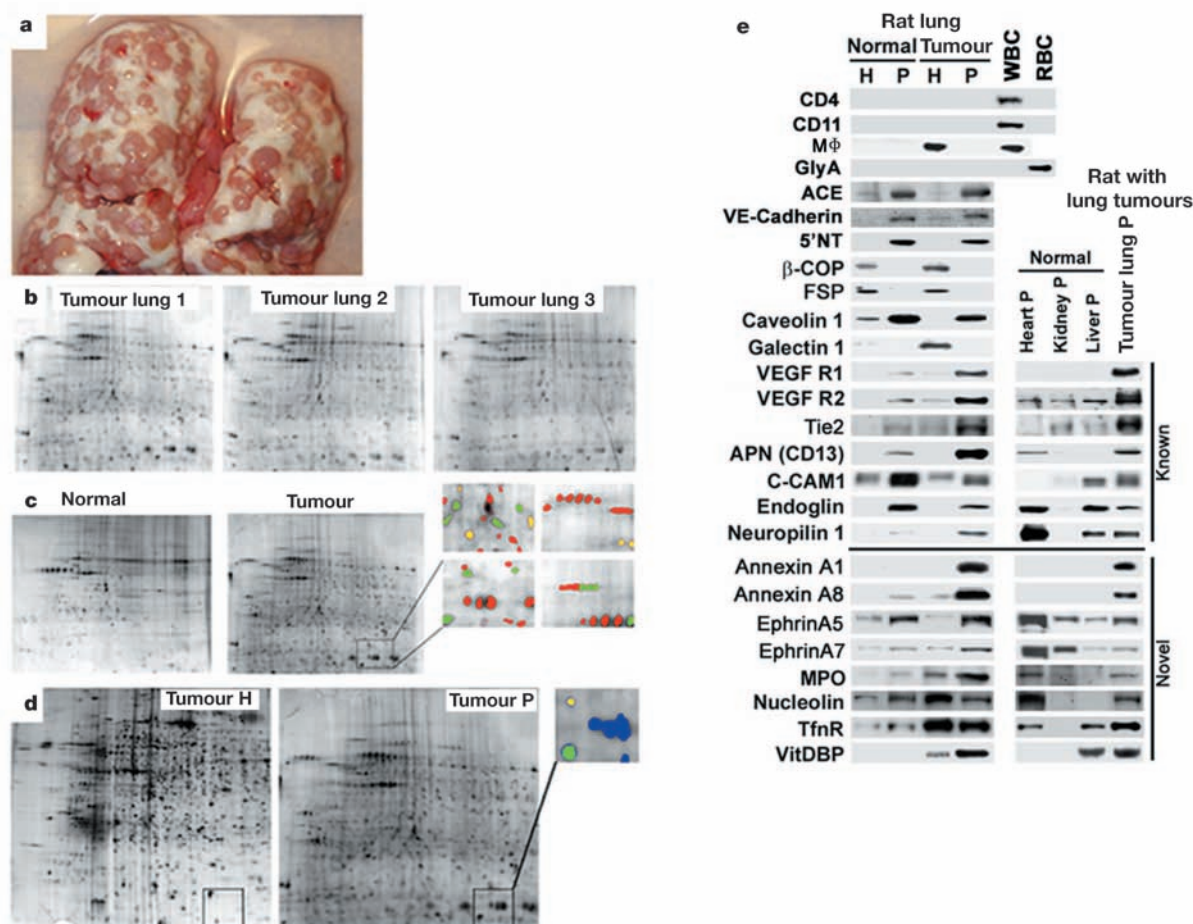


Figure 3 Mapping rat lung tumours and organs for differential endothelial cell surface protein expression *in vivo*. **a**, Image of representative lungs with multiple solid tumours. **b**, 2D gel analysis of silica-coated luminal endothelial plasma membranes isolated from breast adenocarcinoma tumour-bearing lungs of three different animals. **c**, 2D gel analysis of endothelial plasma membranes from normal rat lungs (left) and lungs bearing breast adenocarcinoma tumours (right). A colour difference map of four small gel regions between normal and tumour endothelial plasma membrane proteins shows relative abundance of protein spots. Green, tumour endothelial plasma membranes > normal endothelial plasma membranes. Yellow, tumour endothelial plasma membranes < normal endothelial plasma membranes. Red, only in the tumour endothelial plasma membranes. Red-edged, tumour endothelial plasma

membranes = normal endothelial plasma membranes. **d**, 2D gel analysis of tumour homogenate (H) (left) and endothelial plasma membranes (P) (right). Yellow, protein spots enriched in the homogenate over plasma membranes. Green, protein spots enriched in plasma membrane over homogenate. Blue, detected only in tumour endothelial plasma membranes and not homogenate. **e**, Western blot analysis of proteins (10 µg) from rat lung homogenate (H) and endothelial plasma membranes (P) from normal organs or tumour-bearing rat lungs using antibodies to the indicated proteins. RBC, red blood cell; WBC, white blood cell; MΦ, macrophage; 5' NT, 5' nucleotidase; GlyA, glycophorin A; FSP, fibroblast surface protein; VEGF R, vascular endothelial growth factor receptor; APN, aminopeptidase N; MPO, myeloperoxidase; TfR, transferrin receptor; VitDBP, vitamin D binding protein.



Figure 4 Tumour-specific targeting of ^{125}I -AnnA1 antibodies after intravenous injection into tumour-bearing rats. **a**, Whole-body planar γ -scintigraphic imaging 4 h after tail vein injection of ^{125}I -AnnA1 antibodies ($50\ \mu\text{Ci}$ of IgG at $10\ \text{Ci g}^{-1}$) into lung-tumour-bearing female rats. **b**, ^{125}I -AnnA1 antibody signal superimposed onto photo of experimental animal lying on the detector plate. **c, d**, Tumour-bearing lungs were excised and imaged *ex vivo*. Digital image of excised lungs showing location of tumours, circled in yellow (**c**). Overlay of planar images of tumour hot spots with excised tumour-bearing lungs (**d**).

shown), showing the specific ability of our antibodies to bind APP exposed on the endothelial cell surface of the lung but not other organs.

Consistent with OX-45 expression in various leukocytes³⁸, we detected OX-45 (but not APP) in blood. Thus, imaging with the OX-45 antibody showed little targeting to the lung (data not shown). Biodistribution analysis confirmed only 2% of the injected dose per gram of tissue accumulating in the lung and greater than 70% of the injected dose per gram of tissue remaining in the blood, primarily bound to the buffy coat subfraction of the blood.

APP was reported to be expressed on mouse blood vessels of normal breast and mammary adenocarcinomas as a target for a homing peptide³⁹. High resolution single photon emission computed tomography (SPECT) imaging of ^{125}I -APP antibodies injected intravenously into female rats detected no accumulation in normal breast tissue (Fig. 2a) or in primary or metastatic breast tumour lesions (Fig. 2b; and multiple images taken over 36 hours (data not shown)). SPECT imaging provided a clear three-dimensional view of normal lungs (Fig. 2c) which was in striking contrast to the irregular images, with multiple nodular regions lacking signal, in rats with mammary adenocarcinomas in the lung (Fig. 2d). Thus, based on the immuno-targeting imaged *in vivo*, endothelial APP expression in the rat appeared quite specific for normal lung tissue, apparently induced by conditions present in the normal lung tissue microenvironment, but absent in the tumour

milieu *in vivo*, even when the tumour blood vessels are derived from the normal lung vasculature.

Tumour-induced endothelial cell proteins

To determine whether the tumour microenvironment in the lung is sufficiently different to induce new endothelial protein expression, we isolated silica-coated luminal endothelial cell plasma membranes and their caveolae from normal rat lungs and lungs bearing breast adenocarcinomas^{35,36} (Fig. 3a). As above, these isolates were significantly enriched relative to the tissue homogenates in endothelial cell surface markers (caveolin, 5' nucleotidase, ACE and VE-cadherin) and markedly depleted in markers of possible contaminants (including β -COP, CD4, CD11, glycophorin A, fibroblast surface protein and galectin-1, that is expressed by the tumour cells⁴⁰) (Fig. 3e and data not shown). The endothelial plasma membranes isolated from tumours expressed multiple angiogenesis markers (enriched relative to both tumour homogenates and normal lung endothelial cell plasma membranes). Lastly, markers of immune cells known to infiltrate solid tumours were detected in tumour homogenates but not its endothelial cell plasma membranes.

We used 2D gels to visualize several hundred protein spots in endothelial cell plasma membranes from normal lungs versus tumours in lungs. These maps were reproducible (Fig. 3b). Multiple protein spots were detected in these tumour but not normal endothelial plasma membranes (Fig. 3c). Prominent 2D spots easily detected in tumour plasma membranes were not detected in the homogenates (Fig. 3d), consistent with the small percentage of endothelial cell plasma membranes in the tumours. Tissue subfractionation appeared necessary to unmask differentially expressed tumour vascular proteins obscured by the molecular complexity of the total tumour.

Again, we applied a subtractive proteomic approach using antibody and mass spectrometric analysis of these plasma membranes and caveolae isolated from them to identify 15 differentially expressed proteins (Fig. 3e), including proteins already implicated in tumour angiogenesis: vascular endothelial growth factor (VEGF) receptors-1 and -2; Tie2; aminopeptidase-N; endoglin; carcino-embryonic antigen-related cell adhesion molecule 1 (CD66)/ (C-CAM-1) and neuropilin-1^{27,28}. These proteins were enriched in the tumour endothelial plasma membranes relative to tumour homogenates, consistent with proper subfractionation. Eight new tumour-induced vascular proteins were also identified: annexin A1; annexin A8; ephrin A5; ephrin A7; myeloperoxidase; nucleolin; transferrin receptor and vitamin D-binding protein. Consistent with the subtractive screen hypothesis, 12 out of 15 proteins were

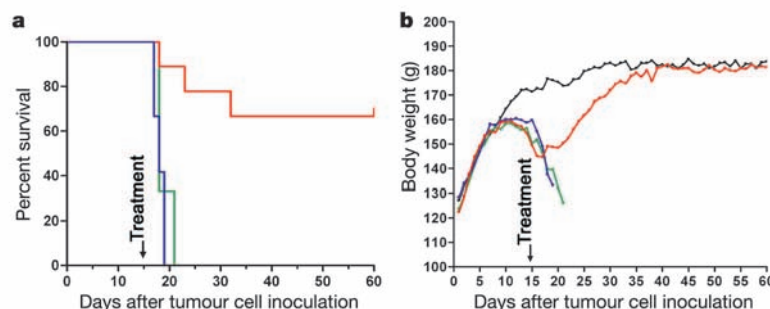


Figure 5 AnnA1 targeted radio-immunotherapy increases survival in rats. Female Fisher rats were injected intravenously with 13762 cells to induce lung tumours (Day 0). **a**, Kaplan-Meier survival curve comparing the survival of tumour-bearing rats injected 15 days after tumour cell inoculation with ^{125}I -AnnA1 antibodies ($50\ \mu\text{g}$ of IgG at $10\ \mu\text{Ci g}^{-1}$, red line; $n = 10$) versus control non-targeting ^{125}I -IgG (blue line; $n = 10$)

versus untreated animals (green line; $n = 3$). **b**, Body weights of rats were measured daily and plotted. Red, tumour-bearing rats treated with ^{125}I -AnnA1 antibodies; blue, tumour-bearing rats treated with control ^{125}I -IgG; green, untreated tumour-bearing animals; black, untreated normal rats. $n = 3$. Maximum standard deviation $<10\ \text{g}$ for each weight plotted.

much more evident in tumour endothelial plasma membranes than in normal plasma membranes.

Expression profiling using endothelial plasma membranes isolated from major organs demonstrated that almost all of these proteins exist at the endothelial cell surface of at least one major organ, albeit mostly at levels much less than tumours (Fig. 3e). One promising tumour candidate target was the 34 KDa protein recognized by annexin A1 (AnnA1) antibodies only in tumour endothelial plasma membrane. Tissue immuno-histochemistry confirmed tumour blood vessel reactivity (data not shown). Thus, the tumour microenvironment appeared to induce distinct protein expression on the endothelial cell surface.

Targeting and imaging of solid tumours

Annexins are cytosolic proteins that can associate with cell membranes in a calcium-dependent manner⁴¹. Some annexins may translocate the lipid bilayer to the external cell surface⁴¹. To test whether AnnA1 is sufficiently exposed and tumour vessel specific to permit immuno-targeting *in vivo*, we performed whole-body imaging using ¹²⁵I-labelled AnnA1 monoclonal antibodies. γ -scintigraphic planar images captured four hours postinjection showed a distinct focus of radioactivity in the lung and little signal elsewhere, in the bodies of rats with tumours (Fig. 4a, b). Non-targeting ¹²⁵I-labelled IgG did not show detectable selective tumour uptake (data not shown). When the lungs were imaged *ex vivo*, we observed ¹²⁵I-AnnA1 antibody accumulation in the tumour as hot spots corresponding to visible tumours (Fig. 4c, d). Furthermore, targeting was prevented by a 30-fold excess of unlabelled AnnA1 IgG but not by control IgG (data not shown). Region-of-interest and biodistribution analysis confirmed targeting *in vivo* with an average tumour accumulation at two hours of 34% injected dose per gram

of tissue. This compared favourably to VEGF receptor antibodies which accumulated at 6.4% injected dose per gram of tumour. When injected into rats without tumours, ¹²⁵I-AnnA1 antibodies showed no targeting of normal organs, including lung at levels less than 1% injected dose per gram of tissue, whereas VEGF receptor antibody accumulation was greater in multiple organs (data not shown). The uptake ratio in rat tumour-bearing lungs versus normal lungs was up to 2.0 and 70 for antibodies to VEGF receptor and AnnA1, respectively. Thus, γ -scintigraphic imaging rapidly validated AnnA1 as a tumour target that is readily accessible to antibody injected intravenously for tumour targeting and imaging *in vivo*. AnnA1 appeared to be selectively externalized on the endothelial cell surface in the solid tumours.

Radio-immunotherapy of solid tumours

Many tumour-bearing rats imaged with the ¹²⁵I-AnnA1 antibody survived, so we performed a survival study and recorded animal body weights; 80% of the animals treated with the ¹²⁵I-AnnA1 antibody survived 8 days or longer compared with the ¹²⁵I-IgG-treated and untreated rats, which died within 7 days (Fig. 5a). The body weights of all tumour-bearing rats began to drop 7–10 days after tumour cell inoculation (Fig. 5b). The control rats continued to decrease in body weight to 23–30% less than normal at their death. In contrast, rats treated with ¹²⁵I-AnnA1 IgG began to gain weight within 3–4 days and reached a normal body weight after 25 days. This increased survival was striking because in this model many animals die within 2–4 days of treatment and thus may lack sufficient time to benefit from treatment. The survival rate of rats surviving the first week approached 90%. The one rat that died after 2 weeks required euthanasia because of a leg tumour and large tail tumour that were not apparent when treated. Thus, a single

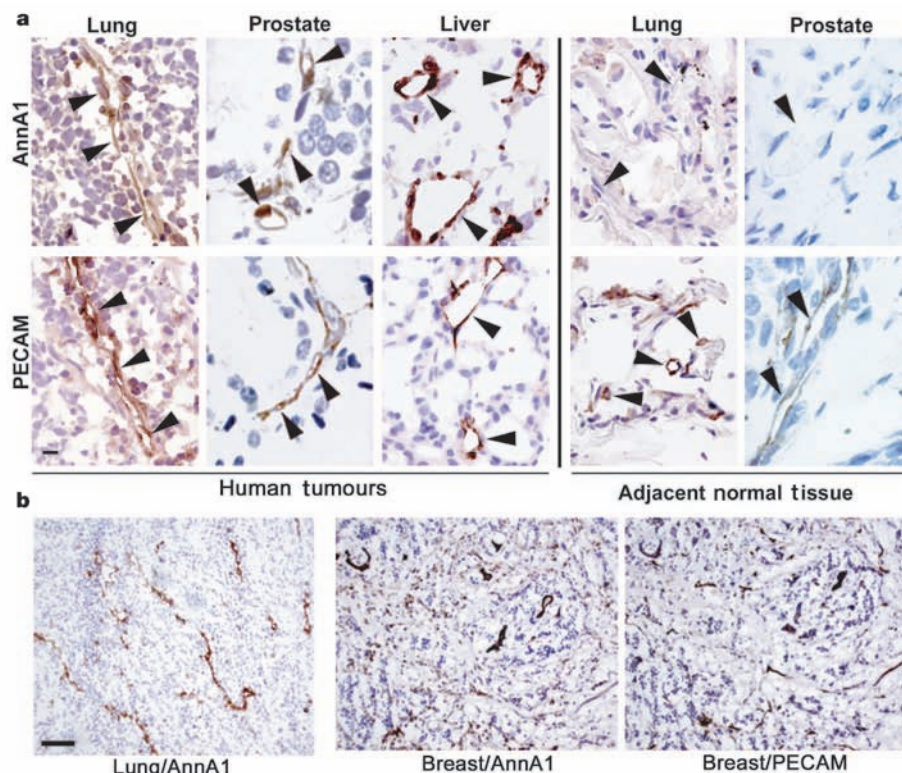


Figure 6 Expression of AnnA1 on vascular endothelium of multiple primary human solid tumours. **a**, Immuno-staining of paraffin-embedded (prostate) or frozen (lung, liver) sections from human tumour specimens and normal tissues, as indicated, with antibodies to AnnA1 (top panels) and PECAM (bottom panels). Arrowheads indicate blood vessels.

Scale bar represents 10 μ m. **b**, Low magnification images of human lung and breast tumour specimens stained with antibodies to AnnA1 or PECAM, as indicated. Scale bar represents 100 μ m.

injection of ^{125}I -AnnA1 antibody caused significant remission even in advanced disease.

AnnA1 in human tumour neovasculature

We immuno-stained tissue sections of human solid tumours. AnnA1 antibody labelled blood vessels of human prostate, liver, breast and lung tumours but not matched normal tissues (Fig. 6). Antibodies to platelet endothelial cell adhesion molecule-1 (PECAM) stained both normal and tumour blood vessels. The lack of AnnA1 expression in vascular endothelium of multiple normal organs has been reported previously^{42–45}. Thus, AnnA1 was selectively detected on the neovascular endothelium of multiple human solid tumours.

Discussion

We describe a new logic-based discovery and validation approach that integrates newly developed, global analytical techniques to map the proteins expressed *in vivo* at the luminal endothelial cell surface. This approach has demonstrated distinct molecular signatures in normal and neoplastic tissues. We applied this strategy to rat lungs and solid tumours to uncover, from the vast number of proteins expressed in tissue, approximately 50 differentially expressed proteins, including two promising tissue-selective endothelial cell surface proteins that permitted rapid and specific immuno-targeting and imaging *in vivo*. Profiling the endothelial cell surface that is accessible *in vivo* is an important logical step for validating the hypothesis of tissue-modulated endothelial cell diversity *in vivo* and for discovering potential targets. This is important not only to direct pharmacodelivery and molecular imaging but also—by means of caveolae—to overcome the restrictive endothelial cell barrier to transport drugs or genes to the underlying tissue cells^{16,17}.

Key features of our analysis included tissue subfractionation with subtractive proteomic and bioinformatic filters that together reduced tissue complexity by greater than five orders of magnitude, and unmasked a manageable subset of proteins at the inherently accessible blood–tissue interface. This has led to the creation of a database of vascular endothelial cell surface proteins containing accessible proteins apparently modulated by the tissue. After bioinformatic filtering, western blot analysis and/or tissue immuno-staining provided sensitive expression profiling to yield only a few candidate proteins exhibiting organ- or tumour-specificity. Planar and SPECT imaging, using intravenously injected antibodies specific to each protein, visualized targeting and tissue accumulation with high sensitivity and resolution. Imaging thus provided a rigorous and objective validation of accessibility and tissue-specificity of the few remaining candidate targets and ultimately demonstrated potential utility such as, how rapidly and to what extent the antibody targets single organs or solid tumours *in vivo*.

So far, *in vivo* imaging correlates well with the expression profiling using isolated endothelial plasma membranes. It showed lung-specific immuno-targeting for APP (Fig. 2), tumour-specificity for AnnA1 (Fig. 4) and targeting of multiple tissues for less specific target proteins (that is, PV-1, VEGF-R and podocalyxin) (our unpublished observations). Because of poor access inside many tissues, antibodies injected intravenously usually require significantly higher doses ($\geq 250 \mu\text{g kg}^{-1}$ (ref. 46) compared with $20 \mu\text{g kg}^{-1}$ used here) to get a small percentage into the tissue that binds and can then be visualized clearly a day or so later, after the background signal has been cleared from the blood. Here, the binding was direct and unhindered by barriers so that both tissue accumulation and blood clearance were rapid, thereby providing striking images within minutes to hours. The rapidity and high levels of specific targeting observed here meet the theoretical expectation of the vascular targeting strategy¹⁷. Although fast emerging as a standard for assessing targeting and specificity^{1–4}, whole-body imaging may be underutilized in target validation despite being non-invasive and highly sensitive.

Using a new, integrated analytical approach we have found new tissue-specific endothelial cell surface proteins that can act as vascular targets to bring us one step closer to achieving the century-old elusive goal of targeting single organs or solid tumours *in vivo*^{1,2,4,6,10}. The development and optimization of several key technologies has provided the sensitivity and throughput required for this new target discovery and validation process. We expect that site-directed vascular and caveolar targeting will benefit both drug and gene delivery in the treatment of many diseases¹⁶. We have demonstrated here, using a rat tumour model that monoclonal antibodies to AnnA1 can effectively direct low levels of radio-nuclides ($100 \mu\text{Ci}$) to concentrate in, and thus destroy solid tumours, which ultimately increases animal survival. Because AnnA1 is also selectively detected in multiple human solid tumours, this target may similarly help to image and treat human disease. □

Methods

A detailed description of the materials, 2D gel analysis, expression profiling *in vivo* of candidate proteins, rat tumour models, γ -scintigraphic imaging and biodistribution analysis is available in Supplementary Methods.

Proteomic analysis

We used a three-pronged approach to resolve proteins/peptides for mass spectrometric analysis: (1) The proteins of silica-coated luminal endothelial cell plasma membranes and caveolae isolated from various normal organs or tumours were resolved by high resolution 2D gel analysis before excising potential organ- or tumour-specific protein spots for mass spectrometric analysis; (2) Proteins in the isolated endothelial plasma membranes and caveolae were separated using one-dimensional (1D) SDS–polyacrylamide gel electrophoresis (SDS–PAGE) and specific bands of interest were cut out or the whole gel lane was cut into 50 slices, and the proteins in each of the gel slices were analysed by mass spectrometry; and (3) Multi-dimensional Protein Identification Technology (MudPIT) was used to analyse tryptic peptides from a complex mixture of proteins extracted directly from the whole isolated endothelial plasma membrane fraction and caveolae isolates. Each gel spot/slice was de-stained and digested overnight with trypsin before extracting the cleaved peptides from the gel and then loading them onto a reverse phase micro-column for gradient acetonitrile elution directly into the mass spectrometer (LCQ Deca XP ion trap mass spectrometer (ThermoFinnigan) equipped with a modified micro-electrospray ionization source from Mass Evolution). For MudPIT, 150 μg of complex peptide mixture was separated by 2D liquid chromatography, comprising a micro-column packed with three phases of chromatographic material as follows: 8.5 cm of 5 μm reversed phase material (Polaris C18-A, Metachem); 4 cm of 5 μm 300 Å strong cation exchanger (PolyLC); and lastly 3.5 cm of C₁₈ material, using a helium pressure cell operated at 600–900 psi (Mass Evolution). Peptides were directly eluted into the mass spectrometer using 2D chromatography with 18 step-elutions from the strong cation exchanger, followed by a gradient elution of the reversed phase material. Operation of the quarterary Agilent 1100 HPLC pump and the mass spectrometer was fully automated during the entire procedure using the Excalibur 1.2 data system (ThermoFinnigan). Continuous cycles of one full scan (m/z 400 to 1,400) followed by three data-dependent MS/MS measurements at 35% normalized collision energy were performed. MS/MS measurements were allowed for the three most intense precursor ions with an enabled exclusion list of 25 m/z values (± 1.5 Da) or a maximum time limit of 5 min. The zoom scan function to determine the charge state was disabled in order to increase the duty cycle of the instrument.

Database search and *in silico* analysis of tandem mass spectra

MS/MS spectra were extracted from raw files requiring a minimum of 21 signals with an intensity of at least 4.75×10^4 arbitrary units (AU). Extracted MS/MS spectra were automatically assigned to the best matching peptide sequence using the SEQUEST algorithm and the Sequest Browser software package (ThermoFinnigan). SEQUEST searches were performed using a rat protein database containing 40,800 protein sequences that were downloaded as FASTA formatted sequences from ENTREZ (National Center for Biotechnology Information (NCBI); <http://www.ncbi.nlm.nih.gov/Entrez>). Sequence redundancies were removed using Perl script. The peptide mass search tolerance was set to 3 Da. Spectral matches were retained with a minimal cross-correlation score (XCorr) of 1.5, 2.2 and 3.3 for charge states +1, +2 and +3, respectively. DeltaCN (top match's XCorr minus the second-best match's XCorr, divided by top match's XCorr) had to be equal or less than 0.07. Retained spectral matches were filtered and re-assigned to proteins using DTASelect. DTASelect outputs of independent measurements were entered into the AVATAR database. AVATAR was designed to store a large amount of mass spectrometric data and to provide tools to analyse the data to extract valuable information. We used relational models for database design based on entity relationship and implemented the database in the MySQL relational database management system (MySQL) to support database query and management. This relational database plus the Perl-based interface greatly improved data organization, data consistency and integrity, and facilitated data comparison and information retrieval. In the case of the 1D gel and MudPIT approaches, AVATAR was used to subtract the data to find proteins detected only at the lung endothelial cell surface *in vivo* versus *in vitro* or on the tumour but not normal endothelium.

In silico bioinformatic interrogation

To identify possible candidates for intravenously accessible targets from the subset of proteins identified as lung or tumour induced, we determined their currently known membrane-associated structure (bilayer spanning versus lipid anchor (intra- or extracellular) versus peripheral interaction) from scientific reports and/or protein databases, such as SwissProt (<http://us.expasy.org/sprot/sprot-top.html>) and the NCBI ENTREZ protein database (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>). We also used web-based prediction programs to identify candidates that may harbour transmembrane spanning alpha helices (Prediction of Transmembrane Regions and Orientation (TMPred), http://www.ch.embnet.org/software/TMPRED_form.html) or glycosylation sites to indicate a possible ectodomain exposed to the circulating blood (Prosit Scan, http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_prosite.html). Only 100% probabilities were taken into consideration.

Received 28 January; accepted 19 April 2004; doi:10.1038/nature02580.

- Massoud, T. F. & Gambhir, S. S. Molecular imaging in living subjects: seeing fundamental biological processes in a new light. *Genes Dev.* **17**, 545–580 (2003).
- Herschman, H. R. Molecular imaging: looking at problems, seeing solutions. *Science* **302**, 605–608 (2003).
- Rudin, M. & Weissleder, R. Molecular imaging in drug discovery and development. *Nature Rev. Drug Discov.* **2**, 123–131 (2003).
- Weissleder, R. Scaling down imaging: molecular mapping of cancer in mice. *Nature Rev. Cancer* **2**, 11–18 (2002).
- Drews, J. Drug discovery: a historical perspective. *Science* **287**, 1960–1964 (2000).
- Lindsay, M. A. Target discovery. *Nature Rev. Drug Discov.* **2**, 831–838 (2003).
- Workman, P. New drug targets for genomic cancer therapy: successes, limitations, opportunities and future challenges. *Curr. Cancer Drug Targets* **1**, 33–47 (2001).
- Anzick, S. L. & Trent, J. M. Role of genomics in identifying new targets for cancer therapy. *Oncology (Huntingt.)* **16**, 7–13 (2002).
- Cavenee, W. K. Genetics and new approaches to cancer therapy. *Carcinogenesis* **23**, 683–686 (2002).
- Huber, L. A. Is proteomics heading in the wrong direction? *Nature Rev. Mol. Cell Biol.* **4**, 74–80 (2003).
- Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
- Jain, R. K. The next frontier of molecular medicine: delivery of therapeutics. *Nature Med.* **4**, 655–657 (1998).
- Dvorak, H. F., Nagy, J. A. & Dvorak, A. M. Structure of solid tumors and their vasculature: implications for therapy with monoclonal antibodies. *Cancer Cells* **3**, 77–85 (1991).
- von Mehren, M., Adams, G. P. & Weiner, L. M. Monoclonal antibody therapy for cancer. *Annu. Rev. Med.* **54**, 343–369 (2003).
- Farah, R. A., Clinchy, B., Herrera, L. & Vitetta, E. S. The development of monoclonal antibodies for the therapy of cancer. *Crit. Rev. Eukaryot. Gene Expr.* **8**, 321–356 (1998).
- Carver, L. A. & Schnitzer, J. E. Caveolae: mining little caves for new cancer targets. *Nature Rev. Cancer* **3**, 571–581 (2003).
- Schnitzer, J. E. Vascular targeting as a strategy for cancer therapy. *N. Engl. J. Med.* **339**, 472–474 (1998).
- Madri, J. A. & Williams, S. K. Capillary endothelial cell culture: Phenotype modulation by matrix components. *J. Cell Biol.* **97**, 153–165 (1983).
- Schnitzer, J. E. in *Vascular Endothelium: Physiology, Pathology and Therapeutic Opportunities* (eds Born, G. V. R. & Schwartz, C. J.) 77–95 (Schattauer, Stuttgart, 1997).
- Aird, W. C. *et al.* Vascular bed-specific expression of an endothelial cell gene is programmed by the tissue microenvironment. *J. Cell Biol.* **138**, 1117–1124 (1997).
- Janzer, R. C. & Raff, M. C. Astrocytes induce blood–brain barrier properties in endothelial cells. *Nature* **325**, 253–257 (1987).
- Stewart, P. A. & Wiley, M. J. Developing nervous tissue induces formation of blood–brain barrier characteristics in invading endothelial cells: a study using quail–chick transplantation chimeras. *Dev. Biol.* **84**, 183–192 (1981).
- Auerbach, R. *et al.* Specificity of adhesion between murine tumor cells and capillary endothelium: an *in vitro* correlate of preferential metastasis *in vivo*. *Cancer Res.* **47**, 1492–1496 (1987).
- Pasqualini, R. & Ruoslahti, E. Organ targeting *in vivo* using phage display peptide libraries. *Nature* **380**, 364–366 (1996).
- Rajotte, D. *et al.* Molecular heterogeneity of the vascular endothelium revealed by *in vivo* phage

display. *J. Clin. Invest.* **102**, 430–437 (1998).

- Papetti, M. & Herman, I. M. Mechanisms of normal and tumor-derived angiogenesis. *Am. J. Physiol. Cell Physiol.* **282**, C947–C970 (2002).
- Ferrara, N. VEGF and the quest for tumour angiogenesis factors. *Nature Rev. Cancer* **2**, 795–803 (2002).
- Kerbel, R. & Folkman, J. Clinical translation of angiogenesis inhibitors. *Nature Rev. Cancer* **2**, 727–739 (2002).
- St Croix, B. *et al.* Genes expressed in human tumor endothelium. *Science* **289**, 1197–1202 (2000).
- Huang, X. *et al.* Tumor infarction in mice by antibody-directed targeting of tissue factor to tumor vasculature. *Science* **275**, 547–550 (1997).
- McIntosh, D. P., Tan, X.-Y., Oh, P. & Schnitzer, J. E. Targeting endothelium and its dynamic caveolae for tissue-specific transcytosis *in vivo*: A pathway to overcome cell barriers to drug and gene delivery. *Proc. Natl Acad. Sci. USA* **99**, 1996–2001 (2002).
- Sipkins, D. A. *et al.* Detection of tumor angiogenesis *in vivo* by alphaVbeta3-targeted magnetic resonance imaging. *Nature Med.* **4**, 623–626 (1998).
- Hu, Z. & Garen, A. Targeting tissue factor on tumor vascular endothelial cells and tumor cells for immunotherapy in mouse models of prostatic cancer. *Proc. Natl Acad. Sci. USA* **98**, 12180–12185 (2001).
- Hood, J. D. *et al.* Tumor regression by targeted gene delivery to the neovasculature. *Science* **296**, 2404–2407 (2002).
- Oh, P. & Schnitzer, J. E. in *Cell Biology: A Laboratory Handbook* (ed. Celis, J.) 34–36 (Academic Press, Orlando, 1998).
- Schnitzer, J. E., McIntosh, D. P., Dvorak, A. M., Liu, J. & Oh, P. Separation of caveolae from associated microdomains of GPI-anchored proteins. *Science* **269**, 1435–1439 (1995).
- Schnitzer, J. E. gp60 is an albumin-binding glycoprotein expressed by continuous endothelium involved in albumin transcytosis. *Am. J. Physiol.* **262**, H246–H254 (1992).
- Henniker, A. J., Bradstock, K. F., Grimsley, P. & Atkinson, M. K. A novel non-lineage antigen on human leucocytes: characterization with two CD-48 monoclonal antibodies. *Dis. Markers* **8**, 179–190 (1990).
- Essler, M. & Ruoslahti, E. Molecular specialization of breast vasculature: a breast-homing phage-displayed peptide binds to aminopeptidase P in breast vasculature. *Proc. Natl Acad. Sci. USA* **99**, 2252–2257 (2002).
- Perillo, N. L., Marcus, M. E. & Baum, L. G. Galectins: versatile modulators of cell adhesion, cell proliferation, and cell death. *J. Mol. Med.* **76**, 402–412 (1998).
- Gerke, V. & Moss, S. E. Annexins: from structure to function. *Physiol. Rev.* **82**, 331–371 (2002).
- Dreier, R., Schmid, K. W., Gerke, V. & Riehemann, K. Differential expression of annexins I, II and IV in human tissues: an immunohistochemical study. *Histochem. Cell Biol.* **110**, 137–148 (1998).
- Eberhard, D. A., Brown, M. D. & VandenBerg, S. R. Alterations of annexin expression in pathological neuronal and glial reactions. Immunohistochemical localization of annexins I, II (p36 and p11 subunits), IV, and VI in the human hippocampus. *Am. J. Pathol.* **145**, 640–649 (1994).
- Ahn, S. H., Sawada, H., Ro, J. Y. & Nicolson, G. L. Differential expression of annexin I in human mammary ductal epithelial cells in normal and benign and malignant breast tissues. *Clin. Exp. Metastasis* **15**, 151–156 (1997).
- McKanna, J. A. & Zhang, M. Z. Immunohistochemical localization of lipocortin 1 in rat brain is sensitive to pH, freezing, and dehydration. *J. Histochem. Cytochem.* **45**, 527–538 (1997).
- Bredow, S., Lewin, M., Hofmann, B., Marecos, E. & Weissleder, R. Imaging of tumour neovasculature by targeting the TGF-beta binding receptor endoglin. *Eur. J. Cancer* **36**, 675–681 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Wempen, K. Hearn, M. Bourne, L. Randall and T. Smith for technical assistance. This research was supported by the National Institute of Health (Heart, Lung and Blood), National Cancer Institute, Sidney Kimmel, Schutz Foundation, California Tobacco-related Disease Research Program and California Breast Cancer Research Program.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.E.S. (jschnitzer@skcc.org).

The interstellar N₂ abundance towards HD 124314 from far-ultraviolet observations

David C. Knauth, B-G Andersson, Stephan R. McCandliss & H. Warren Moos

Department of Physics and Astronomy, The Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218, USA

The abundance of interstellar molecular nitrogen (N₂) is of considerable importance: models of steady-state gas-phase interstellar chemistry^{1,2}, together with millimetre-wavelength observations^{3,4} of interstellar N₂H⁺ in dense molecular clouds predict that N₂ should be the most abundant nitrogen-bearing molecule in the interstellar medium. Previous attempts to detect N₂ absorption in the far-ultraviolet⁵ or infrared⁶ (ice features) have hitherto been unsuccessful. Here we report the detection of interstellar N₂ at far-ultraviolet wavelengths towards the moderately reddened star HD 124314 in the constellation of Centaurus. The N₂ column density is larger than expected from models of diffuse clouds and significantly smaller than expected for dense molecular clouds¹. Moreover, the N₂ abundance does not explain the observed variations⁷ in the abundance of atomic nitrogen (N I) towards high-column-density sightlines, implying that the models of nitrogen chemistry in the interstellar medium are incomplete⁸.

Although N₂ was unsuccessfully searched for with the Copernicus satellite⁵, the Far Ultraviolet Spectroscopic Explorer (FUSE) enables sensitive probes of sightlines with large amounts of interstellar extinction. Archival observations of HD 124314 were acquired on 22 March 2000 using the large aperture (30" × 30") in the histogram data collection mode. HD 124314 is an O6Vnf star⁹ with visual magnitude $V = 6.64$ mag (ref. 10) and a moderate amount of interstellar reddening ($E(B-V) = 0.53$; ref. 11) equivalent to 1.5 magnitudes of visual extinction (A_V) that is located at a distance $d = 700^{+1,500}_{-240}$ pc from the Sun. (ref. 12). The data cover the wavelength range 905–1,185 Å with a spectral resolution of about 20 km s⁻¹. FUSE^{13,14} is particularly well suited for studies of interstellar N₂, because all electronic ground-state transitions are only observable below 1,000 Å. The instrument generally provides two independent spectra for each wavelength region, which can be cross-checked. For the present study, we focus on the two silicon carbide (SiC) detectors covering wavelengths below 1,000 Å.

The data were reduced and calibrated with version 2.4.1 of the CalFUSE pipeline (ref. 15), which incorporates the appropriate Doppler corrections to remove the effects of spacecraft motion and place the data on the heliocentric velocity scale. Owing to the lack of onboard wavelength calibration, a zero-point uncertainty remains in the dispersion solution. To minimize uncertainties (± 6 km s⁻¹ per pixel; ref. 16) caused by the relative wavelength calibration, we used a cross-correlation technique to combine the individual exposures. To fix the absolute heliocentric velocity scale (V_{helio}), we used interstellar H₂ lines in the FUSE data and archival CO data from the Space Telescope Imaging Spectrograph (STIS) on board the Hubble Space Telescope (HST), assuming CO and H₂ to be coexistent¹⁷. The final summed SiC 1B (SiC 2A) spectrum has a signal-to-noise ratio of 45 (50) per 9-pixel-resolution element at 958 Å.

Figure 1 shows the normalized 957–961 Å portion of the FUSE spectrum of HD 124314. Several strong absorption lines due to interstellar molecular hydrogen (H₂), deuterated molecular hydrogen (HD), and N I (refs 18–20) are clearly detected superposed on the broad stellar continuum, revealing at least two, barely resolved,

interstellar velocity components. In addition, there are weaker absorption features found in both SiC detector segments at the expected location of the 0–0 band of the $c'_4^1\Sigma_u^+ - X^1\Sigma_g^+$ transition of N₂ at 958.6 Å with a band oscillator strength (f_{band}) of 0.145 (ref. 21). These features are much too narrow to be stellar in nature because the star has a rotational velocity of 300 km s⁻¹ (ref. 22). Examination of other, less reddened O6 stars, see Supplementary Fig. 1, reveal no evidence for photospheric absorption at this wavelength. No detector defects are present in either detector and other than N₂, no interstellar species have identified features at this wavelength.

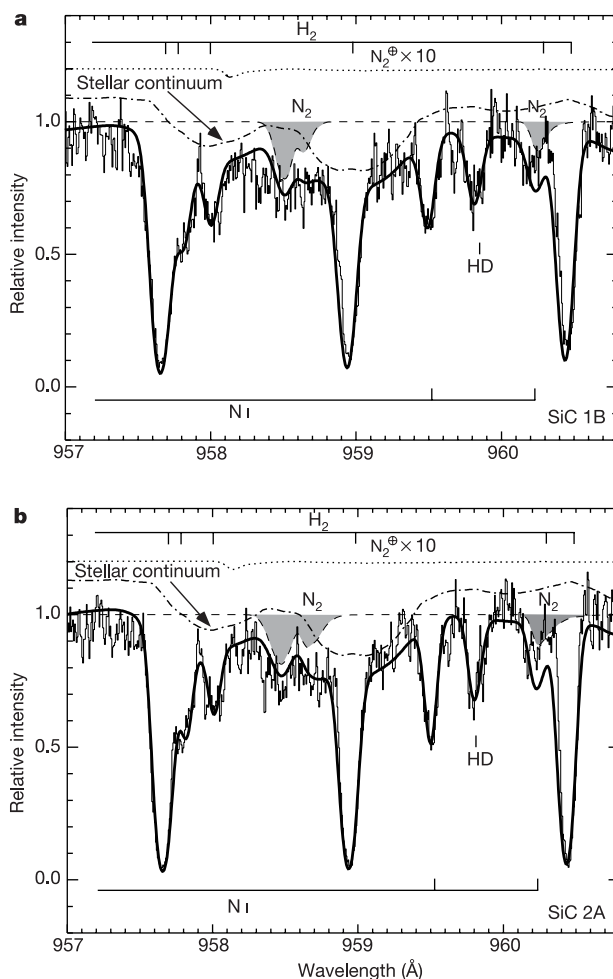


Figure 1 The N₂ portion of the normalized FUSE spectra of HD 124314. **a**, The SiC 1B data and **b**, the SiC 2A data (thin histograms). The thick solid line is the best fit of the data for the wavelength range 951–968 Å. Note that only the range from 957–961 Å is displayed here. All obvious interstellar absorption features (for example, H₂, N I, and so on) in this region are marked. The dot-dashed line is a stellar continuum model²³ with estimated stellar parameters of $T_{\text{eff}} = 37,500$ K and $\log(g) = 3.5$. The stellar He II line at 959 Å line changes dramatically with the stellar parameters—and this directly affects the longer-wavelength N₂ trough at 958 Å. Inclusion of the c_3 N₂ band at 960 Å is vital in constraining the N₂ abundance. Therefore, the choice of stellar model does not change our detection of interstellar N₂. Before being incorporated into the profile-fitting program, the stellar model was rotationally broadened to 300 km s⁻¹ (ref. 22). The offset between the data and the stellar model is caused solely by the line wing opacity (τ) of H₂. The best fit required two velocity components separated by 14 km s⁻¹ for H₂, HD and N I, and similar results are obtained when using the O I velocity separation. Unrealistic parameter values resulted for a two-component fit to N₂. However, limiting the fit to one component in N₂, at the higher velocity, produced convergent fits. The dashed line filled with shading represents the normalized N₂ absorption profiles ($e^{-\tau}$). Our detection of interstellar N₂ is not contaminated by telluric N₂ (dotted line), which has been scaled by a factor of ten and shifted up to 1.2 for the sake of clarity.

Many of the other N₂ bands are undetected owing to severe blending with more abundant species (for example, H I and H₂). For completeness, we have included the 0–0 band of the c₃¹Σ_u⁺ – X¹Σ_g⁺ transition of N₂ at 960.3 Å (*f*_{band} = 0.056) in our study because individual oscillator strengths (*f*-values) are available (G. Stark, personal communication), and although this band is blended with absorption from N I at 960.201 Å and H₂ P(5) at 960.271 Å, the blending is not as severe as for the other stronger N₂ bands. Additionally, there are two N₂ bands outside the fitted wavelength range that are relatively unblended, but are so weak (*f*_{band} ≈ 0.02) that they are not detectable.

The N₂ absorption spectrum of the c₄' and c₃ transitions consists of many rotational lines²¹ under interstellar conditions. Our profile synthesis code⁷ explicitly treats the N₂ excitation and uses the laboratory wavelengths and *f*-values²¹ for the individual rotational lines, thereby reproducing the true N₂ line profiles. To evaluate whether the observed spectral features are due to interstellar N₂, we independently compared the 958 Å region of the SiC 1B and SiC 2A data to a synthetic spectrum combining a stellar model²³ and interstellar medium absorption from H₂ (*J* = 0–5), HD, N I, and N₂. The total synthetic spectrum was convolved to the FUSE resolution and then fitted to the data using a χ² minimization technique. No attempt was made to fit the weak CO band at 956 Å (ref. 24) because it does not significantly affect our fit. The profile synthesis parameters—temperature (*T*), Doppler broadening parameter (*b*-value) and column density (*N*)—were adjusted until a minimum χ² was obtained. Our best fits are shown overlying the data in Fig. 1 and the results are listed in Table 1. The apparent slight mismatch of some features in Fig. 1 (for example, the N I line at 960.2 Å) arise because the fit is based on a larger spectral region (951–968 Å) than is shown. Two velocity components are required to fit the interstellar H₂, HD, and N I, whereas only one component is required for N₂. It is important to note that because of the low instrumental resolution, we cannot reliably constrain the *b*-value for N₂. For numerical stability, the *b*-values of all species in the fit were constrained to coincide. Given that N₂ is likely to be restricted to the densest gas along the line of sight, its *b*-value might be smaller than that found for the other species. Therefore, our derived N₂ column must be treated as a lower limit.

Another consideration that needs to be addressed is the presence, if any, of telluric N₂ absorption. FUSE observations are designed to stop science exposures when the limb angle (θ_L) reaches 15°. There are a small number of exposures in the FUSE archives that extend below this, but the data quality is too low to place useful limits on the contribution expected from terrestrial N₂. The Hopkins Ultraviolet Telescope detected telluric N₂ absorption towards only those observations that went below θ_L < 15° (P. Feldman, personal communication). All our observations were taken with θ_L > 20°, so we calculated the expected contribution of telluric N₂ (that is, *N*(N₂)_{telluric}) using the National Space Science Data Center MSIS

model of the Earth's atmosphere²⁵ for the date of our observation, the satellite's altitude of 750 km, and θ_L = 20°. At this altitude, the space density of telluric N₂ is 2,200 cm^{−3} with a characteristic temperature of 1,100 K. We find *N*(N₂)_{telluric} = 1.1 × 10¹² cm^{−2} for θ_L = 20°. As shown in Fig. 1, telluric N₂ with these characteristics does not significantly contribute to the observed spectrum.

To further assess our N₂ detection, we used ultra-high-resolution archival O I, S I, and CO STIS/HST data. HST observed HD 124314 on 10 April 1999 at a velocity resolution of 1.5 km s^{−1}. The normalized spectra are shown in Fig. 2. Because CO is a tracer of gas shielded from dissociating ultraviolet radiation^{17,26,27}, the facts that

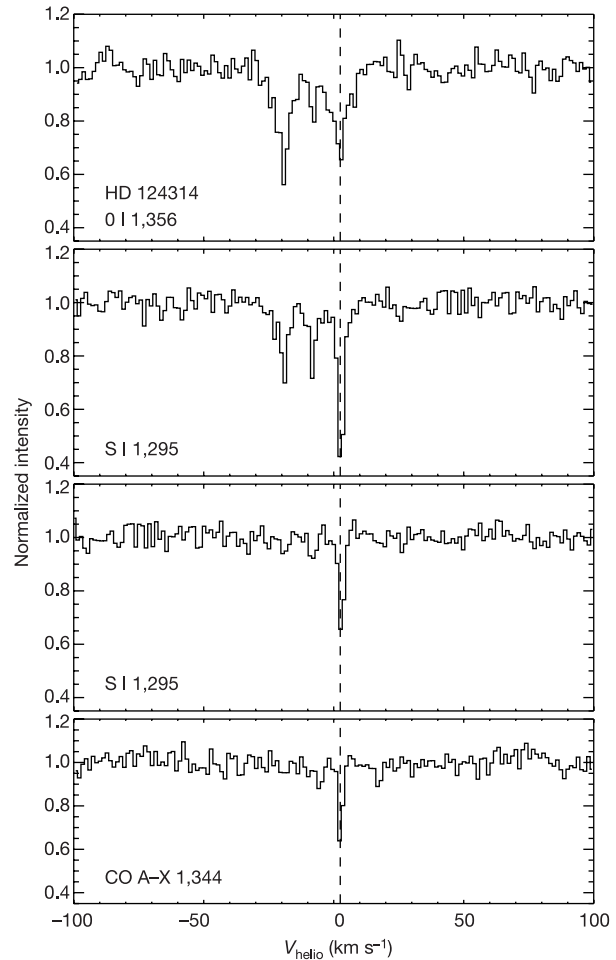


Figure 2 The normalized archival STIS/HST spectra of O I, S I and CO. This data reveals at least three velocity components in Interstellar O I line at 1,356 Å (two are dominant) and S I line at 1,295 Å, but only one component in Interstellar S I line at 1,296 Å and CO bands at 1,344 Å, 1,322 Å and 1,301 Å. These data indicate that the CO component is the densest interstellar cloud along the line of sight. Using a curve-of-growth analysis, the measured CO column density is well determined as: *N*(CO) = (7.6 ± 1.1) × 10¹³ cm^{−2}. We cannot accurately derive the *b*-value from these optically thin lines. We find that the S I doublet is slightly saturated. A curve of growth calculated for these two data points yields a S I column density of *N*(S I) = (1.6 ± 0.2) × 10¹³ cm^{−2} and a *b*-value of 1.0 ± 0.1 km s^{−1} for the longer-wavelength component. The velocity separation of 22 km s^{−1} between the two dominant O I components is similar (within the uncertainties of the FUSE data) to the velocity separation of 14 km s^{−1} we find from our analysis of H₂. The dashed line at ~3 km s^{−1} represents the velocity to which the longer-wavelength H₂ component was normalized. The agreement in velocity separation between the two measured H₂ components and the two strongest O I components provides strong support for the adopted velocity structure used in our model. The fact that both N₂ and CO are found in the single densest component provides further support for the first detection of interstellar N₂.

Table 1 Results from profile synthesis

Species	<i>N</i> _{SiC 2A}	Literature values ^{7,10}
<i>N</i> (H ₂) (component 1)	(1.5 ± 0.2) × 10 ²⁰ cm ^{−2}	...
<i>N</i> (H ₂) (component 2)	(1.4 ± 0.2) × 10 ²⁰ cm ^{−2}	(3.3 ± 0.4) × 10 ²⁰ cm ^{−2}
<i>T</i> _{1–0} (component 1)	340 ± 100 K	...
<i>T</i> _{1–0} (component 2)	38 ± 3 K	54 K
<i>N</i> (N I) (component 1)	(8.5 ± 1.0) × 10 ¹⁵ cm ^{−2}	...
<i>N</i> (N I) (component 2)	(1.9 ± 0.3) × 10 ¹⁷ cm ^{−2}	(2.21 ± 0.27) × 10 ¹⁷ cm ^{−2}
<i>N</i> (N ₂) (component 2)	(4.6 ± 0.8) × 10 ¹³ cm ^{−2*}	...
<i>T</i> _{N₂} (component 2)	36 ± 5 K	...
<i>V</i> _{helioc} (component 1)	3.0 ± 0.5 km s ^{−1†}	...
<i>V</i> _{helioc} (component 2)	−17.0 ± 0.5 km s ^{−1}	...
<i>b</i> -value (component 1)	6.0 ± 0.5 km s ^{−1}	...
<i>b</i> -value (component 2)	7.0 ± 0.5 km s ^{−1}	...
Reduced χ ²	2.47	...

Results for SiC 1B agree within their mutual uncertainties to those for SiC 2A reported here.

* Average of SiC 1B and SiC 2A results.

† Set to coincide with the STIS/HST heliocentric velocity scale.

N_2 and CO both occur in the same velocity component, that the velocity separation in O I and H_2 agree, and that the CO component is aligned in velocity with the higher-velocity O I component strongly supports our claim of the first detection of interstellar N_2 . Our N_2 detection is further supported by the fact that the excitation temperatures of H_2 (1–0) and of N_2 agree for the longer-wavelength component. Although the N_2 b -value is not reliably determined, we use the SI b -value to improve our estimate and find $N(N_2) = (4.6 \pm 0.8) \times 10^{13} \text{ cm}^{-2}$. We note that SI may be more extended than CO or N_2 in interstellar space. Because nitrogen has a lower cosmic abundance compared to either carbon or oxygen, we believe that $N(N_2)$ should not be larger than $N(\text{CO})$ and should the true N_2 b -value be much smaller than that for SI (although this is unlikely), significant amounts of N_2 (that is, $N(N_2) > 10^{14} \text{ cm}^{-2}$) could be present. Our firm lower limit to $N(N_2)$ is not sensitive to the b -value or the choice of stellar model used to represent the stellar continuum. Taken together, these data strongly indicate that N_2 has been detected for the first time in the interstellar medium and with a column density of $N(N_2) > 3.8 \times 10^{13} \text{ cm}^{-2}$.

From our analysis of the O I line at $1,356 \text{ \AA}$, we find $N(\text{O I}) = (6.77 \pm 0.50) \times 10^{17} \text{ cm}^{-2}$ for the component near 3 km s^{-1} . Using the observed $N(\text{O I})/N(H_{\text{tot}})$ ratio $(4.74 \pm 0.81) \times 10^{-4}$ (ref. 11), this component contains $N(H_{\text{tot}}) = 1.5 \times 10^{21} \text{ cm}^{-2}$, where $N(H_{\text{tot}}) = N(\text{H I}) + 2N(H_2)$. The amount of interstellar reddening for this component can be determined from the dust-to-gas ratio²⁸, which yields $E(B-V) = 0.26 \text{ mag}^{-1}$. Assuming that the ratio of total to selective extinction is 3.1 (ref. 29), we find that the total visual extinction is 0.8 mag. When we compare our results to models of interstellar gas-phase chemistry¹, none of the standard cloud models explain our observations. The observed N_2 fractional abundance is more than two orders of magnitude too low for dense cloud models and approximately two orders of magnitude larger than expected from models of diffuse clouds. The fact that N I shows a deficiency in its relative abundance for lines of sight with $N(H_{\text{tot}}) > 10^{21} \text{ cm}^{-2}$ would argue that dense cloud chemistry should be important for interstellar nitrogen. However, the measured N_2 abundance and upper limits for other sightlines⁷ do not account for the observed variations. Additionally, we find that the fractional abundance of N_2 towards HD 124314 is $N_2/H_2 = 3.3 \times 10^{-7}$, similar to those estimated from N_2H^+ observations⁴ of dark molecular clouds. Therefore, the far-ultraviolet lines of N_2 provide a unique probe of interstellar nitrogen chemistry in the transition region from diffuse to dense molecular gas. □

Received 16 February; accepted 30 April 2004; doi:10.1038/nature02614.

- Viala, Y. P. Chemical equilibrium from diffuse to dense interstellar clouds. I Galactic molecular clouds. *Astron. Astrophys. (Suppl.)* **64**, 391–437 (1986).
- Bergin, E. A., Langer, W. D. & Goldsmith, P. F. Gas-phase chemistry in dense interstellar clouds including grain surface molecular depletion and desorption. *Astrophys. J.* **441**, 222–243 (1995).
- Womack, M., Zirui, L. M. & Wyckoff, S. A survey of N_2H^+ in dense clouds: Implications for interstellar nitrogen and ion-molecule chemistry. *Astrophys. J.* **387**, 417–429 (1992).
- Womack, M., Zirui, L. M. & Wyckoff, S. Estimates of N_2 abundances in dense molecular clouds. *Astrophys. J.* **393**, 188–192 (1992).
- Lutz, B. L., Owen, T. & Snow, T. P. Jr A search with *Copernicus* for interstellar N_2 in diffuse clouds. *Astrophys. J.* **227**, 159–162 (1979).
- Sandford, S. A., Bernstein, M. P., Allamandola, L. J., Goorvitch, D. & Teixeira, T. C. V. S. The abundances of solid N_2 and gaseous CO_2 in interstellar dense molecular clouds. *Astrophys. J.* **548**, 836–851 (2001).
- Knauth, D. C., Andersson, B.-G., McCandless, S. R. & Moos, H. W. Potential variations in the interstellar N I abundance. *Astrophys. J.* **596**, L51–L54 (2003).
- Le Petit, F., Roueff, E. & Herbst, E. H_3^+ and other species in the diffuse cloud towards ζ Persei: A new detailed model. *Astron. Astrophys.* **417**, 993–1002 (2004).
- Walborn, N. R. The space distribution of the O stars in the solar neighborhood. *Astron. J.* **78**, 1067–1073 (1973).
- Cruz-Gonzalez, C., Recillas-Cruz, E., Costero, R., Peimbert, M. & Torres-Peimbert, S. A catalogue of galactic O stars. The ionization of the low density interstellar medium by runaway stars. *Rev. Mex. Astron. Astrofis.* **1**, 211–259 (1974).
- André, M. K. et al. Oxygen gas-phase abundance revisited. *Astrophys. J.* **591**, 1000–1012 (2003).
- Perryman, M. A. C. et al. The HIPPARCOS catalogue. *Astron. Astrophys.* **323**, L49–L52 (1997).
- Moos, H. W. et al. Overview of the Far Ultraviolet Spectroscopic Explorer mission. *Astrophys. J.* **538**, L1–L6 (2000).
- Sahnow, D. J. et al. On-orbit performance of the Far Ultraviolet Spectroscopic Explorer satellite. *Astrophys. J.* **538**, L7–L11 (2000).

- Dixon, W. V. & Sahnow, D. J. CalFUSE v2.2: An improved data calibration pipeline for the Far Ultraviolet Spectroscopic Explorer (FUSE) in astronomical data analysis software and systems XII. (Spec. Iss.; eds Payne, H.E., Jedrzejewski, R.I. & Hook, R. N.) *ASP Conf. Ser.* **295**, (2003) 241–244.
- Knauth, D. C., Howk, J. C., Sembach, K. R., Lauroesch, J. T. & Meyer, D. M. On the origin of the high-ionization intermediate velocity gas toward HD 14434. *Astrophys. J.* **592**, 964–974 (2003).
- Federman, S. R., Glassgold, A. E., Jenkins, E. B. & Shaya, E. J. The abundance of CO in diffuse interstellar clouds. An ultraviolet survey. *Astrophys. J.* **242**, 545–559 (1980).
- Abgrall, H. & Roueff, E. Wavelengths, oscillator strengths and transition probabilities of the H_2 molecule for Lyman and Werner systems. *Astron. Astrophys. (Suppl.)* **79**, 313–328 (1989).
- Abgrall, H., Roueff, E. & Viala, Y. Vibration-rotation transition probabilities for the ground electronic $X^1\Sigma^+$ state of HD. *Astron. Astrophys. (Suppl.)* **50**, 505–522 (1982).
- Morton, D. C. Atomic data for resonance absorption lines. III. Wavelengths longward of the Lyman limit for the elements hydrogen to gallium. *Astrophys. J.* **242**, 205–238 (2003).
- Stark, G. et al. Line oscillator strength measurements in the 0–0 band of the $c'_4^1\Sigma_u^+ - X^1\Sigma_g^+$ transition of N_2 . *Astrophys. J.* **531**, 321–328 (2000).
- Conti, P. S. & Ebbets, D. Spectroscopic studies of O-type stars. VII. Rotational velocities V_{ini} and evidence for macroturbulent motions. *Astrophys. J.* **213**, 438–447 (1977).
- Lanz, T. & Hubeny, T. A grid of non-LTE line-blanketed model atmospheres of O-type stars. *Astrophys. J. (Suppl.)* **146**, 417–441 (2003).
- Sheffer, Y., Federman, S. R. & Andersson, B.-G. FUSE measurements of Rydberg bands of interstellar CO between 925 and 1150 Å. *Astrophys. J.* **597**, L29–L32 (2003).
- Hedin, A. E. Extension of the MSIS thermospheric model into the middle and lower atmosphere. *J. Geophys. Res.* **96**, 1159–1172 (1991).
- Clayton, R. N. Self shielding in the Solar Nebula. *Nature* **415**, 860–861 (2002).
- Van Dishoeck, E. F. & Black, J. H. The photodissociation and chemistry of interstellar CO. *Astrophys. J.* **334**, 771–802 (1988).
- Bohlin, R. C., Savage, B. D. & Drake, J. F. A survey of interstellar H I from L_α absorption measurements. II. *Astrophys. J.* **242**, 132–142 (1978).
- Snedden, C., Gehr, R. D., Hackwell, J. A., York, D. G. & Snow, T. P. Infrared colors and the diffuse interstellar bands. *Astrophys. J.* **223**, 168–179 (1978).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Fullerton, S. R. Federman, P. Feldman, E. B. Jenkins, P. Sonnentrucker and P. Wannier for discussions and G. Stark for sharing his unpublished N_2 f -values. This work is based on observations made with the NASA-CNES-CSA Far Ultraviolet Spectroscopic Explorer, which is operated for NASA by Johns Hopkins University.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.C.K. (dknauth@pha.jhu.edu).

A precision measurement of the mass of the top quark

DØ Collaboration*

*A list of authors and their affiliations appear at the end of the paper

The standard model of particle physics contains parameters—such as particle masses—whose origins are still unknown and which cannot be predicted, but whose values are constrained through their interactions. In particular, the masses of the top quark (M_t) and W boson (M_W)¹ constrain the mass of the long-hypothesized, but thus far not observed, Higgs boson. A precise measurement of M_t can therefore indicate where to look for the Higgs, and indeed whether the hypothesis of a standard model Higgs is consistent with experimental data. As top quarks are produced in pairs and decay in only about 10^{-24} s into various final states, reconstructing their masses from their decay products is very challenging. Here we report a technique that extracts more information from each top-quark event and yields a greatly improved precision (of $\pm 5.3 \text{ GeV}/c^2$) when compared to previous measurements². When our new result is combined with our published measurement in a complementary decay mode³ and with the only other measurements available², the new world average for M_t becomes⁴ $178.0 \pm 4.3 \text{ GeV}/c^2$. As a

result, the most likely Higgs mass increases from the experimentally excluded⁵ value⁶ of 96 to 117 GeV/ c^2 , which is beyond current experimental sensitivity. The upper limit on the Higgs mass at the 95% confidence level is raised from 219 to 251 GeV/ c^2 .

The discovery of the top quark in 1995 served as one of the major confirmations of the validity of the standard model (SM)^{7,8}. Of its many parameters, the mass of the top quark, in particular, reflects some of the most crucial aspects of the SM. This is because, in principle, the top quark is point-like and should be massless; yet, through its interactions with the hypothesized Higgs field, the physical mass of the top quark appears to be about the mass of a gold nucleus. Because it is so heavy, the top quark (along with the W boson) provides an unusually sensitive tool for investigating the Higgs field. M_W is known to a precision of 0.05%, while the uncertainty on M_t is at the 3% level¹. Improvements in both measurements are required to restrict further the allowed range of mass for the Higgs; however, given the large uncertainty in M_t , an improvement in its precision is particularly important. As has been pointed out recently^{9,10}, a potential problem for the SM is that, on the basis of the currently accepted value for M_t , the most likely value of the Higgs mass⁶ lies in a range that has already been excluded by experiment⁵. Precise knowledge of the Higgs mass is crucial for our understanding of the SM and any possible new physics beyond it. For example, in a large class of supersymmetric models (theoretically preferred solutions to the deficiencies of the SM), the Higgs mass has to be less than about 135 GeV/ c^2 . Although, unlike the SM, supersymmetry predicts more than one Higgs boson, the properties of the lightest one are expected to be essentially the same as those for the SM Higgs boson. Thus, if the SM-like Higgs is heavier than about 135 GeV/ c^2 , it would disfavour a large class of supersymmetric models. In addition, some of the current limits on supersymmetric particles from LEP¹¹ are extremely sensitive to M_t . In fact, for M_t greater than 179 GeV/ c^2 , the bounds on one of the major supersymmetry parameters, $\tan\beta$, which relates the properties of the SM-like Higgs boson and its heavier partners, would disappear completely¹². Hence, in addition to the impact on searches for the Higgs boson, other important consequences call for improved precision on M_t , and this goal is the main subject of this paper.

The DØ experiment at the Fermilab Tevatron has studied a sample of $t\bar{t}$ events produced in proton-antiproton ($p\bar{p}$) interactions¹³. The total energy of 1.8 TeV released in a head-on collision of a 900-GeV p and a 900-GeV $\bar{p}$ is almost as large as the rest energy of ten gold nuclei. Each top (antitop) quark decays almost immediately into a bottom $b(\bar{b})$ quark and a W^+ (W^-) boson, and we have reexamined those events in which one of the W bosons decays into a charged lepton (electron or muon) and a neutrino, and the other W into a quark and an antiquark (see Fig. 1). These events and their selection criteria are identical to those used to extract the mass of the

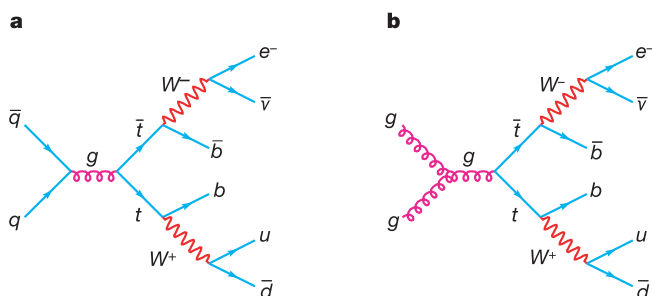


Figure 1 Feynman diagrams for $t\bar{t}$ production in $p\bar{p}$ collisions, with subsequent decays into an electron, neutrino, and quarks. Quark-antiquark production (a) is dominant, but gluon fusion (b) contributes $\sim 10\%$ to the cross-section. This particular final state ($e\bar{\nu}u\bar{b}b\bar{d}$) is one of the channels used in the analysis.

top quark in our previous publication, and correspond to an integrated luminosity of 125 events per pb. (That is, given the production cross-section of the $t\bar{t}$ in $p\bar{p}$ collisions at 1.8 TeV of 5.7 pb, as measured by DØ¹⁴, these data correspond to approximately 700 produced $t\bar{t}$ pairs, a fraction of which is fully detected in various possible decay modes. Approximately 30% of these correspond to the lepton + jets topology categorized in Fig. 2, where 'jet' refers to products of the fragmentation of a quark into a collimated group of particles that are emitted along the quark's original direction.) The main background processes correspond to multijet production (20%), where one of the jets is reconstructed incorrectly as a lepton, and the W + jets production with leptonic W decays (80%), which has the same topology as the $t\bar{t}$ signal.

The previous DØ measurement of M_t in this lepton + jets channel is $M_t = 173.3 \pm 5.6$ (stat) ± 5.5 (syst) GeV/ c^2 , and is based on 91 candidate events. Information pertaining to the older analysis and the DØ detector can be found elsewhere^{13,15}.

The new method of M_t measurement is similar to one suggested previously (ref. 16 and references therein, and ref. 17) for $t\bar{t}$ dilepton decay channels (where both W bosons decay leptonically), and used in previous mass analyses of dilepton events³, and akin to an approach suggested for the measurement of the mass of the W boson at LEP^{18–20}. The critical differences from previous analyses in the lepton + jets decay channel lie in: (1) the assignment of more weight to events that are well measured or more likely to correspond to $t\bar{t}$ signal, and (2) the handling of the combinations of final-state objects (lepton, jets and imbalance in transverse momentum, the latter being a signature for an undetected neutrino) and their identification with top-quark decay products in an event (such as from ambiguity in choosing jets that correspond to b or $\bar{b}$ quarks from the decays of the t and $\bar{t}$ quarks). Also, because leading-order matrix elements were used to calculate the event weights, only events with exactly four jets are kept in this analysis, resulting in a candidate sample of 71 events. Although we are left with fewer events, the new method for extracting M_t provides substantial improvement in both statistical and systematic uncertainties.

We calculate as a function of M_t the differential probability that the measured variables in any event correspond to signal. The maximum of the product of these individual event probabilities provides the best estimate of M_t in the data sample. The impact of biases from imperfections in the detector and event-reconstruction algorithms is taken into account in two ways. Geometric acceptance, trigger efficiencies, event selection, and so on enter through a multiplicative acceptance function that is independent of M_t . Because the angular directions of all the objects in the event, as

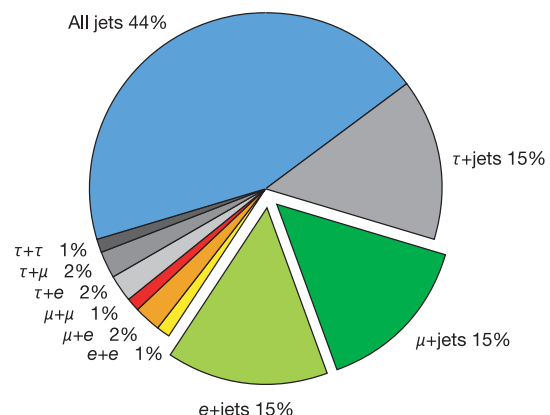


Figure 2 Relative importance of various $t\bar{t}$ decay modes. The 'lepton + jets' channel used in this analysis corresponds to the two offset slices of the pie-chart and amounts to 30% of all the $t\bar{t}$ decays.

well as the electron momentum, are measured with high precision, their measured values are used directly in the calculation of the probability that any event corresponds to $t\bar{t}$ or background production. The known momentum resolution is used to account for uncertainties in measurements of jet energies and muon momenta.

As in the previous analysis¹³, momentum conservation in $\gamma + \text{jet}$ events is used to check that the energies of jets in the experiment agree with Monte Carlo (MC) simulations. This calibration has an uncertainty $\delta E = (0.025E + 0.5 \text{ GeV})$. Consequently, all jet energies in our sample are rescaled by $\pm \delta E$, the analysis redone, and half of the difference in the two rescaled results for M_t ($\delta M_t = 3.3 \text{ GeV}/c^2$) is taken as a systematic uncertainty from this source. All other contributions to systematic uncertainty: MC modelling of signal ($\delta M_t = 1.1 \text{ GeV}/c^2$) and background ($\delta M_t = 1.0 \text{ GeV}/c^2$), effect of calorimeter noise and event pile-up ($\delta M_t = 1.3 \text{ GeV}/c^2$), and other corrections from M_t extraction ($\delta M_t = 0.6 \text{ GeV}/c^2$) are much smaller, and discussed in detail elsewhere^{21,22}. It should be noted that the new mass measurement method provides a significant (about 40%, from ± 5.5 to $\pm 3.9 \text{ GeV}/c^2$) reduction in systematic uncertainty, which is ultimately dominated by the measurement of jet energies. For details on the new analysis, see the Methods.

The final result is $M_t = 180.1 \pm 3.6$ (stat) ± 3.9 (syst) GeV/c^2 . The improvement in statistical uncertainty over our previous measurement is equivalent to collecting a factor of 2.4 as much data. Combining the statistical and systematic uncertainties in quadrature, we obtain $M_t = 180.1 \pm 5.3 \text{ GeV}/c^2$, which is consistent with our previous measurement in the same channel (at about 1.4 standard deviations), and has a precision comparable to all previous M_t measurements combined¹.

The new measurement can be combined with that obtained for the dilepton sample that was also collected at DØ during run I (ref. 3) ($M_t = 168.4 \pm 12.3$ (stat) ± 3.6 (syst) GeV/c^2), to yield the new DØ average for the mass of the top quark:

$$M_t = 179.0 \pm 5.1 \text{ GeV}/c^2 \quad (1)$$

Combining this with measurements from the CDF experiment² provides a new 'world average' (based on all measurements avail-

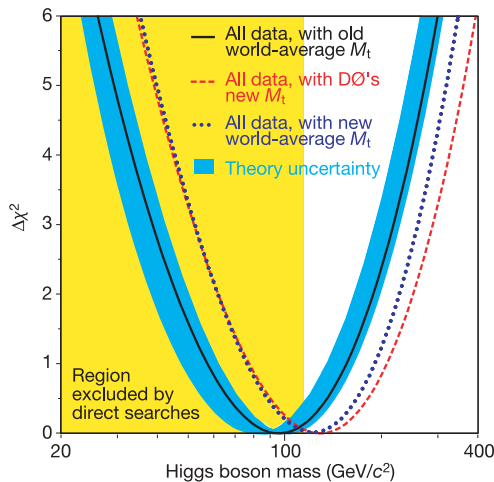


Figure 3 Current experimental constraints on the mass of the Higgs boson. The χ^2 for a global fit to electroweak data⁶ is shown as a function of the Higgs mass. The solid line corresponds to the result for the previous world-averaged $M_t = 174.3 \pm 5.1 \text{ GeV}/c^2$, with the blue band indicating the impact of theoretical uncertainty. The dotted line shows the result for the new world-averaged M_t of $178.0 \pm 4.3 \text{ GeV}/c^2$, whereas the dashed line corresponds to using only the new DØ average of $179.0 \pm 5.1 \text{ GeV}/c^2$. The yellow-shaded area on the left indicates the region of Higgs masses excluded by experiment ($> 114.4 \text{ GeV}/c^2$ at the 95% confidence level¹⁴). The improved M_t measurement shifts the most likely value of the Higgs mass above the experimentally excluded range.

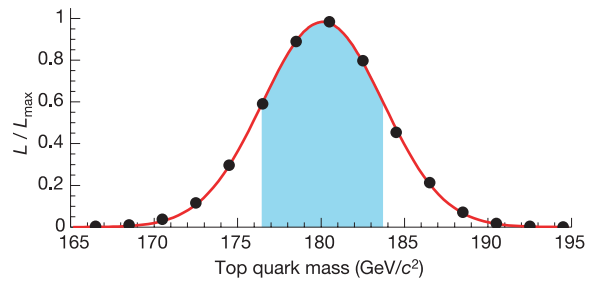


Figure 4 Determination of the mass of the top quark using the maximum-likelihood method. The points represent the likelihood of the fit used to extract M_t divided by its maximum value, as a function of M_t (after a correction for a $-0.5 \text{ GeV}/c^2$ mass bias, see text). The solid line shows a gaussian fit to the points. The maximum likelihood corresponds to a mass of $180.1 \text{ GeV}/c^2$, which is the new DØ measurement of M_t in the lepton + jets channel. The shaded band corresponds to the range of ± 1 standard deviation, and indicates the $\pm 3.6 \text{ GeV}/c^2$ statistical uncertainty of the fit.

able) for the mass of the top quark⁴:

$$M_t = 178.0 \pm 4.3 \text{ GeV}/c^2 \quad (2)$$

dominated by our new measurement. This new world average shifts the best-fit value of the expected Higgs mass from $96 \text{ GeV}/c^2$ to $117 \text{ GeV}/c^2$ (see Fig. 3), which is now outside the experimentally excluded region, yet accessible in the current run of the Tevatron and at future runs at the Large Hadron Collider (LHC), at present under construction at CERN. (The upper limit on the Higgs mass at the 95% confidence level changes from $219 \text{ GeV}/c^2$ to $251 \text{ GeV}/c^2$.) Figure 3 shows the effect of using only the new DØ top mass for fits to the Higgs mass, and indicates a best value of $123 \text{ GeV}/c^2$ and the upper limit of $277 \text{ GeV}/c^2$ at the 95% confidence level. It should be noted that the horizontal scale in Fig. 3 is logarithmic, and the limits on the Higgs boson mass are therefore asymmetric.

The new method is already being applied to data being collected by the CDF and DØ experiments at the new run of the Fermilab Tevatron and should provide even higher precision on the determination of M_t , equivalent to more than a doubling of the data sample, relative to using the conventional method. An ultimate precision of about $2 \text{ GeV}/c^2$ on the mass of the top quark is expected to be reached in several years of Tevatron operation. Further improvement may eventually come from the LHC. □

Methods

The probability density as a function of M_t can be written as a convolution of the calculable cross-section and any effects from measurement resolution:

$$P(x, M_t) = \frac{1}{\sigma(M_t)} \int d\sigma(y, M_t) dq_1 dq_2 f(q_1) f(q_2) W(y, x) \quad (3)$$

where $W(y, x)$, our general transfer function, is the normalized probability for the measured set of variables x to arise from a set of nascent (partonic) variables y , $d\sigma(y, M_t)$ is the partonic theoretical differential cross-section, $f(q)$ are parton distribution functions that reflect the probability of finding any specific interacting quark (antiquark) with momentum q within the proton (antiproton), and $\sigma(M_t)$ is the total cross-section for producing $t\bar{t}$. The integral in equation (3) sums over all possible parton states, leading to what is observed in the detector.

The acceptance of the detector is given in terms of a function $A(x)$ that relates the probability $P_m(x, M_t)$ of measuring the observed variables x to their production probability $P(x, M_t)$: $P_m(x, M_t) = A(x)P(x, M_t)$. Effects from energy resolution, and so on are taken into account in the transfer function $W(y, x)$. The integrations in equation (3) over the eleven well-measured variables (three components of charged-lepton momentum and eight jet angles) and the four equations of energy-momentum conservation leave five integrals that must be performed to obtain the probability that any event represents $t\bar{t}$ (or background) production for some specified value of M_t .

The probability for a $t\bar{t}$ interpretation can be written as:

$$P_{t\bar{t}} = \frac{1}{12\sigma_{t\bar{t}}} \int d^5\Omega \sum_{\text{perm}, \nu} |\mathcal{M}_{t\bar{t}}|^2 \frac{f(q_1)f(q_2)}{|q_1||q_2|} \Phi_6 W_{\text{jets}}(E_{\text{part}}, E_{\text{jet}})$$

where Ω represents a set of five integration variables, $\mathcal{M}_{t\bar{t}}$ is the leading-order matrix element for $t\bar{t}$ production^{23,24}, $f(q_1)$ and $f(q_2)$ are the CTEQ4M parton distribution functions for the incident quarks²⁵, Φ_6 is the phase-space factor for the six-object final state, and the sum is over all 12 permutations of the jets and all possible neutrino solutions.

$W_{\text{jets}}(E_{\text{part}}, E_{\text{jet}})$ corresponds to a function that maps parton-level energies E_{part} to energies measured in the detector E_{jet} and is based on MC studies. A similar expression, using a matrix element for $W + \text{jets}$ production (the dominant background source) that is independent of M_t , is used to calculate the probability for a background interpretation, P_{bkg} .

Studies of samples of HERWIG (ref. 26; we used version 5.1) MC events indicate that the new method is capable of providing almost a factor-of-two reduction in the statistical uncertainty on the extracted M_t . These studies also reveal that there is a systematic shift in the extracted M_t that depends on the amount of background there is in the data. To minimize this effect, a selection is introduced, based on the probability that an event represents background. The specific value of the P_{bkg} cut-off is based on MC studies carried out before applying the method to data, and, for $M_t = 175 \text{ GeV}/c^2$, retains 71% of the signal and 30% of the background. A total of 22 data events out of our 71 candidates pass this selection.

The final likelihood as a function of M_t is written as:

$$\ln(L(M_t)) = \sum_{i=1}^N \ln[c_1 P_{\text{tr}}(x_i, M_t) + c_2 P_{\text{bkg}}(x_i)] - N \int A(x) [c_1 P_{\text{tr}}(x, M_t) + c_2 P_{\text{bkg}}(x)] dx$$

The integration is performed using MC methods. The best value of M_t (when L is at its maximum L_{max}) represents the most likely mass of the top quark in the final N -event sample, and the parameters c_i reflect the amounts of signal and background. MC studies show that there is a downward shift of $0.5 \text{ GeV}/c^2$ in the extracted mass, and this correction is applied to the result. Reasonable changes in the cut-off on P_{bkg} do not have a significant impact on M_t .

Figure 4 shows the value of $L(M_t)/L_{\text{max}}$ as a function of M_t for the 22 events that pass all selection criteria, after correction for the $0.5 \text{ GeV}/c^2$ bias in mass. The likelihood is maximized with respect to the parameters c_i at each mass point. The gaussian fit in the figure yields $M_t = 180.1 \text{ GeV}/c^2$, with a statistical uncertainty of $\delta M_t = 3.6 \text{ GeV}/c^2$. The systematic uncertainty, dominated by the measurement of jet energies, as discussed above, amounts to $\delta M_t = 3.9 \text{ GeV}/c^2$. When added in quadrature to the statistical uncertainty from the fit, it yields the overall uncertainty on the new M_t measurement of $\pm 5.3 \text{ GeV}/c^2$.

Received 23 January; accepted 21 April 2004; doi:10.1038/nature02589.

1. Hagiwara, K. *et al.* Review of particle physics. *Phys. Rev. D* **66**, 010001 (2002).
2. Affolder, T. *et al.* (CDF Collaboration). Measurement of the top quark mass with the Collider Detector at Fermilab. *Phys. Rev. D* **63**, 032003 (2001).
3. Abbott, B. *et al.* (DØ Collaboration). Measurement of the top quark mass in the dilepton channel. *Phys. Rev. D* **60**, 052001 (1999).
4. The CDF Collaboration, the DØ Collaboration, and the TEVATRON Electro-Weak Working Group. Combination of CDF and DØ Results on the Top-Quark Mass. Preprint at <http://www.arXiv.org/hep-ex/0404010> (2004).
5. Barate, R. *et al.* (ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, and LEP Working Group for Higgs boson searches). Search for the standard model Higgs boson at LEP. *Phys. Lett. B* **565**, 61–75 (2003).
6. The LEP Collaborations ALEPH, DELPHI, L3, and OPAL, the LEP Electroweak Working Group, and the SLD Heavy Flavour Group. A combination of preliminary electroweak measurements and constraints on the standard model. Preprint at <http://www.arXiv.org/hep-ex/0312023> (2003).
7. Abe, F. *et al.* (CDF Collaboration). Observation of top quark production in $p\bar{p}$ collisions with the Collider Detector at Fermilab. *Phys. Rev. Lett.* **74**, 2626–2631 (1995).
8. Abachi, S. *et al.* (DØ Collaboration). Observation of the top quark. *Phys. Rev. Lett.* **74**, 2632–2637 (1995).
9. Ellis, J. The 115 GeV Higgs odyssey. *Comments Nucl. Part. Phys. A* **2**, 89–103 (2002).
10. Chanowitz, M. S. Electroweak data and the Higgs boson mass: a case for new physics. *Phys. Rev. D* **66**, 073002 (2002).

DØ Collaboration (Participants are listed in alphabetical order.)

V. M. Abazov¹, B. Abbott², A. Abdesselam³, M. Abolins⁴, V. Abramov⁵, B. S. Acharya⁶, D. L. Adams⁷, M. Adams⁸, S. N. Ahmed⁹, G. D. Alexeev¹, A. Alton¹⁰, G. A. Alves¹¹, Y. Arnoud¹², C. Avila¹³, V. V. Babintsev¹⁴, L. Babukhadia¹⁴, T. C. Bacon¹⁵, A. Baden¹⁶, S. Baffioni¹⁷, B. Baldin¹⁸, P. W. Balm¹⁹, S. Banerjee⁶, E. Barberis²⁰, P. Baringer²¹, J. Barreto¹¹, J. F. Bartlett¹⁸, U. Bassler²², D. Bauer²³, A. Bean²¹, F. Beaudette³, M. Begel²⁴, A. Belyaev²⁵, S. B. Beri²⁶, G. Bernardi²², I. Bertram²⁷, A. Besson¹², R. Beuselinck¹⁵, V. A. Bezzubov⁵, P. C. Bhat¹⁸, V. Bhatnagar²⁶, M. Bhattacharjee¹⁴, G. Blazey²⁸, F. Blekman¹⁹, S. Blessing²⁵, A. Boehnlein¹⁸, N. I. Bojko⁵, T. A. Bolton²⁹, F. Borchering¹⁸, K. Bos¹⁹, T. Bose³⁰, A. Brand³¹, G. Briskin³², R. Brock⁴, G. Brooijmans³⁰, A. Bross¹⁸, D. Buchholz³³, M. Buehler³⁴, V. Buescher³⁴, V. S. Burtovoi⁵, J. M. Butler³⁵, F. Canelli²⁴, W. Carvalho³⁶, D. Casey⁴, H. Castilla-Valdez³⁷, D. Chakraborty²⁸, K. M. Chan²⁴, S. V. Chekulaev⁵, D. K. Cho²⁴, S. Choi³⁸, S. Chopra⁷, D. Claes³⁹, A. R. Clark⁴⁰, B. Connolly²⁵, W. E. Cooper¹⁸, D. Coppage²¹, S. Crépé-Renaudin¹², M. A. C. Cummings²⁸, D. Cutts³², H. da Motta¹¹, G. A. Davis²⁴, K. De³¹, S. J. de Jong⁹, M. Demarteau¹⁸, R. Demina²⁴, P. Demine⁴¹, D. Denisov¹⁸, S. P. Denisov⁵, S. Desai¹⁴, H. T. Diehl¹⁸, M. Diesburg¹⁸, S. Doulas²⁰, L. V. Dudko⁴², L. Duflost³, S. R. Dugad⁶, A. Duperrin¹⁷, A. Dyshkant²⁸, D. Edmunds⁴, J. Ellison³⁸, J. T. Eltzroth³¹, V. D. Elvira¹⁸, R. Engelmann¹⁴, S. Eno¹⁶, G. Eppley⁴³, P. Ermolov⁴², O. V. Eroshin⁵, J. Estrada²⁴, H. Evans³⁰, V. N. Evdokimov⁵, T. Ferbel²⁴, F. Filthaut⁹, H. E. Fisk¹⁸, M. Fortner²⁸, H. Fox³³, S. Fu³⁰, S. Fuess¹⁸, E. Gallas¹⁸, A. N. Galyaev⁵, M. Gao³⁰, V. Gavrilov⁴⁴, R. J. Genik II²⁷, K. Genser¹⁸, C. E. Gerber⁸, Y. Gershtein³², G. Ginther²⁴, B. Gómez¹³, P. I. Goncharov⁵, K. Gounder¹⁸, A. Goussiou⁴⁵, P. D. Grannis¹⁴, H. Greenlee¹⁸, Z. D. Greenwood⁴⁶, S. Grinstein⁴⁷, L. Groer³⁰, S. Grünwald⁴⁸, S. N. Gurzhiev⁵, G. Gutierrez¹⁸, P. Gutierrez², N. J. Hadley¹⁶, H. Haggerty¹⁸, S. Hagopian²⁵, V. Hagopian²⁵, R. E. Hall⁴⁹, C. Han¹⁰, S. Hansen¹⁸, J. M. Hauptman⁵⁰, C. Hebert²¹, D. Hedin²⁸, J. M. Heinmiller⁹, A. P. Heinsohn³⁸, U. Heintz²⁵, M. D. Hildreth⁴⁵, R. Hirosky⁵¹, J. D. Hobbs¹⁴, B. Hoeneisen⁵², J. Huang²³, Y. Huang¹⁰, I. Iashvili³⁸, R. Illingworth¹⁵, A. S. Ito¹⁸, M. Jaffré³, S. Jain², R. Jesik¹⁵, K. Johns⁵³, M. Johnson¹⁸, A. Jonckheere¹⁸, H. Jöstlein¹⁸, A. Juste¹⁸, W. Kahl²⁹, S. Kahn⁷, E. Kajfasz¹⁷, A. M. Kalinin¹, D. Karmanov⁴², D. Karmgard⁴⁵, R. Kehoe⁴, S. Kesiosoglou³², A. Khanov²⁴, A. Kharchilava⁴⁵, B. Klima¹⁸, J. M. Kohli²⁶, A. V. Kostitskiy⁵, J. Kotcher⁷, B. Kothari³⁰, A. V. Kozelov⁵, E. A. Kozlovsky⁵, J. Krane⁵⁰, W. R. Krishnaswamy⁶, P. Krivkova⁵⁴, S. Krzywdzinski¹⁸, M. Kubantsev⁵, S. Kuleshov⁴⁴, Y. Kulik¹⁸, S. Kunori¹⁸, A. Kupco⁵⁵, V. E. Kuznetsov³⁸, G. Landsberg³², M. M. Lee²⁵, A. Leflat⁴², F. Lehner¹⁸, C. Leonidopoulos³⁰, J. Li³¹, Q. Z. Li¹⁸, J. G. R. Lima²⁸, D. Lincoln¹⁸, S. L. Linn²⁵, J. Linnemann⁴, R. Lipton¹⁸, A. Lucotte¹², L. Lueking¹⁸, C. Lundstedt³⁹, C. Luo²³, A. K. A. Maciel²⁸, R. J. Madaras⁴⁰, V. L. Malyshev¹, V. Manankov⁴², H. S. Mao⁵⁶, T. Marshall²³, M. I. Martin²⁸, S. E. K. Mattingly³², A. A. Mayorov⁵, R. McCarthy¹⁴, T. McMahon⁵⁷, H. L. Melanson¹⁸, A. Melnitchouk³², A. Merkin⁴², K. W. Merritt¹⁸, C. Miao³², H. Miettinen⁴³, D. Mihalcea²⁸, N. Mokhov¹⁸, N. K. Mondal⁶, H. E. Montgomery¹⁸, R. W. Moore⁴, Y. D. Mutaf¹⁴, E. Nagy¹⁷, M. Narain³⁵, V. S. Narasimham⁶, N. A. Naumann⁹, H. A. Neal¹⁰, J. P. Negret¹³, S. Nelson²⁵, A. Nomerotski¹⁸, T. Nunnemann¹⁸, D. O'Neil⁴, V. Oguri³⁶, N. Oshima¹⁸, P. Padley⁴³, K. Papageorgiou⁸, N. Parashar⁴⁶,

11. The LEP Collaborations ALEPH, DELPHI, L3 and OPAL, the LEP Higgs Working Group. Searches for the neutral Higgs bosons of the MSSM: preliminary combined results using LEP data collected at energies up to 209 GeV. Preprint at <http://www.arXiv.org/hep-ex/0107030> (2001).
12. Degrandi, G., Heinemeyer, S., Hollik, W., Slavich, P. & Weiglein, G. Towards high-precision predictions for the MSSM Higgs sector. *Eur. Phys. J. C* **28**, 133–143 (2003).
13. Abbott, B. *et al.* (DØ Collaboration). Direct measurement of the top quark mass by the DØ collaboration. *Phys. Rev. D* **58**, 052001 (1998).
14. Abazov, V. M. *et al.* (DØ Collaboration). $t\bar{t}$ production cross section in $p\bar{p}$ collisions at $\sqrt{s} = 1.8 \text{ TeV}$. *Phys. Rev. D* **67**, 012004 (2003).
15. Abachi, S. *et al.* (DØ Collaboration). The DØ detector. *Nucl. Instrum. Methods A* **338**, 185–253 (1994).
16. Dalitz, R. H. & Goldstein, G. R. Test of analysis method for top–antitop production and decay events. *Proc. R. Soc. Lond. A* **445**, 2803–2834 (1999).
17. Kondo, K. *et al.* Dynamical likelihood method for reconstruction of events with missing momentum. 3: Analysis of a CDF high p_T $e\mu$ event as $t\bar{t}$ production. *J. Phys. Soc. Jpn* **62**, 1177–1182 (1993).
18. Berends, F. A., Papadopoulos, C. G. & Pittau, R. On the determination of M_W and TGCs in W -pair production using the best measured kinematical variables. *Phys. Lett. B* **417**, 385–389 (1998).
19. Abreu, P. *et al.* (DELPHI Collaboration). Measurement of the W pair cross-section and of the W Mass in e^+e^- interactions at 172 GeV. *Eur. Phys. J. C* **2**, 581–595 (1998).
20. Juste, A. Measurement of the W mass in e^+e^- annihilation PhD thesis, 1–160, Univ. Autònoma de Barcelona (1998).
21. Estrada, J. Maximal use of kinematic information for extracting the top quark mass in single-lepton $t\bar{t}$ events PhD thesis, 1–132, Univ. Rochester (2001).
22. Canelli, F. Helicity of the W boson in single-lepton $t\bar{t}$ events PhD thesis, 1–241, Univ. Rochester (2003).
23. Mahlon, G. & Parke, S. Angular correlations in top quark pair production and decay at hadron colliders. *Phys. Rev. D* **53**, 4886–4896 (1996).
24. Mahlon, G. & Parke, S. Maximizing spin correlations in top quark pair production at the Tevatron. *Phys. Lett. B* **411**, 173–179 (1997).
25. Lai, H. L. *et al.* (CTEQ Collaboration). Global QCD analysis and the CTEQ parton distributions. *Phys. Rev. D* **51**, 4763–4782 (1995).
26. Marchesini, G. *et al.* HERWIG: a Monte Carlo event generator for simulating hadron emission reactions with interfering gluons. *Comput. Phys. Commun.* **67**, 465–508 (1992).

Acknowledgements We are grateful to our colleagues A. Quadt and M. Mulders for reading of the manuscript and comments. We thank the staffs at Fermilab and collaborating institutions, and acknowledge support from the Department of Energy and National Science Foundation (USA), Commissariat à l'Energie Atomique and CNRS/Institut National de Physique Nucléaire et de Physique des Particules (France), Ministry for Science and Technology and Ministry for Atomic Energy (Russia), CAPES, CNPq and FAPERJ (Brazil), Departments of Atomic Energy and Science and Education (India), Colciencias (Colombia), CONACyT (Mexico), Ministry of Education and KOSEF (Korea), CONICET and UBACyT (Argentina), The Foundation for Fundamental Research on Matter (The Netherlands), PPARC (UK), Ministry of Education (Czech Republic), the A. P. Sloan Foundation, and the Research Corporation.

Authors' contributions We wish to note the great number of contributions made by the late Harry Melanson to the DØ experiment, through his steady and inspirational leadership of the physics, reconstruction and algorithm efforts.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J. Estrada (estrada@fnal.gov).

R. Partridge³², N. Parua¹⁴, A. Patwa¹⁴, O. Peters¹⁹, P. Pétroff³, R. Piegaia⁴⁷, B. G. Pope⁴, H. B. Prosper²⁵, S. Protopopescu⁷, M. B. Przybycien^{33*}, J. Qian¹⁰, S. Rajagopalan⁷, P. A. Rapidis¹⁸, N. W. Reay²⁹, S. Reucroft²⁰, M. Ridel³, M. Rijssenbeek¹⁴, F. Rizatdinova²⁹, T. Rockwell⁴, C. Royon⁴¹, P. Rubinov¹⁸, R. Ruchti⁴⁵, B. M. Sabirov¹, G. Sajot¹², A. Santoro³⁶, L. Sawyer⁴⁶, R. D. Schamberger¹⁴, H. Schellman³³, A. Schwartzman⁴⁷, E. Shabalina⁸, R. K. Shivpuri⁵⁸, D. Shpakov²⁰, M. Shupe⁵³, R. A. Sidwell²⁹, V. Simak⁵⁵, V. Sirotenko¹⁸, P. Slattery²⁴, R. P. Smith¹⁸, G. R. Snow³⁹, J. Snow⁵⁷, S. Snyder⁷, J. Solomon⁸, Y. Song³¹, V. Sorin⁴⁷, M. Sosebee³¹, N. Sotnikova⁴², K. Soustruznik⁵⁴, M. Souza¹¹, N. R. Stanton²⁹, G. Steinbrück³⁰, D. Stoker⁵⁹, V. Stolin⁴⁴, A. Stone⁸, D. A. Stoyanova⁵, M. A. Strang³¹, M. Strauss², M. Strovink⁴⁰, L. Stutte¹⁸, A. Sznajder³⁶, M. Talby¹⁷, W. Taylor¹⁴, S. Tentindo-Repond²⁵, T. G. Trippie⁴⁰, A. S. Turcot⁷, P. M. Tuts³⁰, R. Van Kooten²³, V. Vaniev⁵, N. Varelas⁸, F. Villeneuve-Seguirer¹⁷, A. A. Volkov⁵, A. P. Vorobiev⁵, H. D. Wahl²⁵, Z.-M. Wang¹⁴, J. Warchol⁴⁵, G. Watts⁶⁰, M. Wayne⁴⁵, H. Weerts⁴, A. White³¹, D. Whiteson⁴⁰, D. A. Wijngaarden⁹, S. Willis²⁸, S. J. Wimpenny³⁸, J. Womersley¹⁸, D. R. Wood²⁰, Q. Xu¹⁰, R. Yamada¹⁸, T. Yasuda¹⁸, Y. A. Yatsunenko¹, K. Yip⁷, J. Yu³¹, M. Zanabria¹³, X. Zhang², B. Zhou¹⁰, Z. Zhou⁵⁰, M. Zielinski²⁴, D. Zieminska²³, A. Ziemiński²³, V. Zutshi²⁸, E. G. Zverev⁴² & A. Zylberstein⁴¹

Affiliations for participants: 1, Joint Institute for Nuclear Research, P O Box 79, 141980 Dubna, Russia; 2, University of Oklahoma, Department of Physics and Astronomy, Norman, Oklahoma 73019, USA; 3, Laboratoire de l'Accélérateur Linéaire, IN2P3-CNRS, BP 34, Batiment 200, F-91898 Orsay, France; 4, Michigan State University, Department of Physics and Astronomy, East Lansing, Michigan 48824, USA; 5, Institute for High Energy Physics, 142284 Protvino, Russia; 6, Tata Institute of Fundamental Research, School of Natural Sciences, Homi Bhabha Rd, Mumbai 400005, India; 7, Brookhaven National Laboratory, Physics Department, Bldg 510C, Upton, New York 11973, USA; 8, University of Illinois at Chicago, Department of Physics, 845 W. Taylor, Chicago, Illinois 60607, USA; 9, University of Nijmegen/NIKHEF, P O Box 9010, NL-6500 GL Nijmegen, The Netherlands; 10, University of Michigan, Department of Physics, 500 E. University Avenue, Ann Arbor, Michigan 48109, USA; 11, LAFEX, Centro Brasileiro de Pesquisas Físicas, Rua Dr Xavier Sigaud, 150, 22290-180 Rio de Janeiro, Brazil; 12, Laboratoire de Physique Subatomique et de Cosmologie, IN2P3-CNRS, Université de Grenoble 1, 53 Avenue des Martyrs, F-38026 Grenoble, France; 13, Universidad de los Andes, Department de Física, HEP Group, Apartado Aereo 4976, Bogotá, Colombia; 14, State University of New York, Department of Physics and Astronomy, Stony Brook, New York 11794, USA; 15, Imperial College London, Department of Physics, Prince Consort Road, London SW7 2BW, UK; 16, University of Maryland, Department of Physics, College Park, Maryland 20742, USA; 17, CPPM, IN2P3-CNRS, Université de la Méditerranée, 163 Avenue de Luminy, F-13288 Marseille, France; 18, Fermi National Accelerator Laboratory, P O Box 500, Batavia, Illinois 60510, USA; 19, FOM-Institute NIKHEF and University of Amsterdam/NIKHEF, P O Box 41882, 1009 DB Amsterdam, The Netherlands; 20, Northeastern University, Department of Physics, Boston, Massachusetts 02115, USA; 21, University of Kansas, Department of Physics and Astronomy, 1251 Wescoe Hall Drive, Lawrence, Kansas 66045, USA; 22, LPNHE, Universités Paris VI and VII, IN2P3-CNRS, 4 Place Jussieu, Tour 33, F-75252 Paris, France; 23, Indiana University, Department of Physics, 727 E. 3rd St, Bloomington, Indiana 47405, USA; 24, University of Rochester, Department of Physics and Astronomy, Rochester, New York 14627, USA; 25, Florida State University, Department of Physics 4350, Tallahassee, Florida 32306, USA; 26, Panjab University, Department of Physics, Chandigarh 160014, India; 27, Lancaster University, Department of Physics, Lancaster LA1 4YB, United Kingdom; 28, Northern Illinois University, Department of Physics, DeKalb, Illinois 60115, USA; 29, Kansas State University, Department of Physics, Manhattan, Kansas 66506, USA; 30, Columbia University, Department of Physics, 538 W. 120th St, New York, New York 10027, USA; 31, University of Texas, Department of Physics, Box 19059, Arlington, Texas 76019, USA; 32, Brown University, Department of Physics, 182 Hope St, Providence, Rhode Island 02912, USA; 33, Northwestern University, Department of Physics and Astronomy, 2145 Sheridan Road, Evanston, Illinois 60208, USA; 34, Universität Freiburg, Physikalisches Institut, Hermann-Herder-Strasse 3, 79104 Freiburg, Germany; 35, Boston University, Department of Physics, 590 Commonwealth Avenue, Boston, Massachusetts 02215, USA; 36, Universidade do Estado do Rio de Janeiro, Instituto de Física, Rua São Francisco Xavier, 524, 20559-900 Rio de Janeiro, Brazil; 37, CINVESTAV, Departamento de Física, P O Box 14-740, 07000 Mexico City, Mexico; 38, University of California, Department of Physics, Riverside, California 92521, USA; 39, University of Nebraska, Department of Physics and Astronomy, Lincoln, Nebraska 68588, USA; 40, Lawrence Berkeley National Laboratory and University of California, 1 Cyclotron Road, Berkeley, California 94720, USA; 41, DAPNIA/Service de Physique des Particules, CEA, Saclay, F-91191 Gif-sur-Yvette, France; 42, Moscow State University, Department of Physics, Vorobjovy Gory, 119899 Moscow, Russia; 43, Rice University, Bonner Nuclear Lab, P O Box 1892, Houston, Texas 77005, USA; 44, Institute for Theoretical and Experimental Physics, B. Cheremushkinskaya ul. 25, 117259 Moscow, Russia; 45, University of Notre Dame, Department of Physics, Notre Dame, Indiana 46556, USA; 46, Louisiana Tech University, Department of Physics, Ruston, Louisiana 71272, USA; 47, Universidad de Buenos Aires, Departamento de Física, FCEN, Pabellón 1, Ciudad Universitaria, 1428 Buenos Aires, Argentina; 48, University College Dublin, Department of Experimental Physics, Faculty of Science, Belfield, Dublin 4, Ireland; 49, California State University, Department of Physics, 2345 E. San Ramon Avenue, Fresno, California 93740, USA; 50, Iowa State University, Department of Physics, High Energy Physics Group, Ames, Iowa 50011, USA; 51, University of Virginia, Department of Physics, Charlottesville, Virginia 22901, USA; 52, Universidad San Francisco de Quito, P O Box 17-12-841, Quito, Ecuador; 53, University of Arizona, Department of Physics, P O Box 210081, Tucson, Arizona 85721, USA; 54, Institute of Particle and Nuclear Physics, Center for Particle Physics, Faculty of Mathematics and Physics, Charles University in Prague, V Holesovickach 2, CZ-18000 Prague 8, Czech Republic; 55, Institute of Physics of the Academy of Sciences of the Czech Republic, Center for Particle Physics, Na Slovance 2, CZ-18221 Prague 8, Czech Republic; 56, Institute of High Energy Physics, P O Box 918, Beijing 100039, China; 57, Langston University, Department of Mathematics, Langston, Oklahoma 73050, USA; 58, Delhi University, Department of Physics and Astrophysics, Delhi 110007, India; 59, University of California, Department of Physics and Astronomy, 4129 Frederick Reines Hall, Irvine, California 92697, USA; 60, University of Washington, Department of Physics, P O Box 351560, Seattle, Washington 98195, USA.

*Present addresses: University of Zurich, Zurich, Switzerland (E.L.); Institute of Nuclear Physics, Krakow, Poland (M.B.P.)

Energy-transfer pumping of semiconductor nanocrystals using an epitaxial quantum well

Marc Achermann¹, Melissa A. Petruska¹, Simon Kos¹, Darryl L. Smith¹, Daniel D. Koleske² & Victor I. Klimov¹

¹Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

²Sandia National Laboratories, Albuquerque, New Mexico 87185, USA

As a result of quantum-confinement effects, the emission colour of semiconductor nanocrystals can be modified dramatically by simply changing their size^{1,2}. Such spectral tunability, together with large photoluminescence quantum yields and high photostability, make nanocrystals attractive for use in a variety of

light-emitting technologies—for example, displays, fluorescence tagging³, solid-state lighting and lasers⁴. An important limitation for such applications, however, is the difficulty of achieving electrical pumping, largely due to the presence of an insulating organic capping layer on the nanocrystals. Here, we describe an approach for indirect injection of electron–hole pairs (the electron–hole radiative recombination gives rise to light emission) into nanocrystals by non-contact, non-radiative energy transfer from a proximal quantum well that can in principle be pumped either electrically or optically. Our theoretical and experimental results indicate that this transfer is fast enough to compete with electron–hole recombination in the quantum well, and results in greater than 50 per cent energy-transfer efficiencies in the tested structures. Furthermore, the measured energy-transfer rates are sufficiently large to provide pumping in the stimulated emission regime, indicating the feasibility of nanocrystal-based optical amplifiers and lasers based on this approach.

Several programs worldwide (for example, ref. 5) emphasize the need for efficient solid-state emitters in applications ranging from displays and traffic signs to solid-state lighting. Semiconductor nanocrystals are considered to be promising nanoscale colour-selectable emitters that combine high (potentially 100%) photoluminescence quantum yields with chemical flexibility and processability. Even in the form of a single monolayer they can produce significant power outputs of the order of watts per cm^2 (estimated for a nanocrystal packing density of 10^{12} cm^{-2} , a radiative lifetime of 20 ns, and a moderate quantum yield of 20%).

One approach to nanocrystal-based, electrically pumped light-emitting devices uses hybrid organic/inorganic structures, in which the charges are delivered to nanocrystals through the organic network and/or percolated nanocrystal subsystem^{6–8}. The performance of these devices is, however, limited by low carrier mobilities in both the organic and nanocrystal components, and by the poor stability of organic molecules with respect to photooxidation. Here, we explore a novel, ‘non-contact’ approach to carrier injection into nanocrystals using Förster-like non-radiative energy transfer from a proximal epitaxial quantum well. As energy transfer relies on Coulomb interactions rather than a direct wavefunction overlap, it is not significantly inhibited by the nanocrystal capping layer, and it can potentially allow for an efficient energy flow from a quantum well to nanocrystals. In a real-life device, the quantum well can be pumped electrically in the same way a common quantum-well light-emitting diode is pumped. However, in this work we intentionally

use pulsed optical excitation to study the dynamics of energy transfer as well as the dynamics of other competing processes such as carrier recombination in the quantum well.

The structure studied in our experiments is depicted in Fig. 1a. It consists of an InGa_N quantum well, on top of which is assembled a close-packed monolayer of highly monodisperse CdSe/ZnS core/shell nanocrystals using the Langmuir–Blodgett technique. The nanocrystals were synthesized as described in refs 2 and 9 and consist of a CdSe core (radius = 1.9 nm) overcoated with a shell of ZnS (~0.6 nm thickness), followed by a final layer of the organic molecules trioctylphosphine (TOP) and trioctylphosphine oxide (TOPO). These nanocrystals show efficient emission centred near 575 nm and a structured absorption spectrum with the lowest 1S absorption maximum at ~560 nm (Fig. 1b). Quantum-well samples were grown on sapphire substrates by metal–organic chemical-vapour deposition¹⁰. They consisted of a 20-nm GaN nucleation layer, a 3- μm GaN bottom barrier, and a 3-nm InGa_N quantum well that was either terminated with a 3-nm GaN top barrier (capped quantum well) or remained uncapped. The concentration of In in the quantum wells was 5–10%, which corresponds to an emission wavelength of ~400 nm (Fig. 1b). This wavelength is in the range of strong nanocrystal absorption, which provides strong coupling of quantum-well excitations to the absorption dipole of nanocrystals, and should allow efficient energy transfer. To study energy-transfer dynamics, we monitored the temporal evolution of photoluminescence in the quantum well and the nanocrystals using a time-correlated single-photon counting system that provides ~30 ps time resolution. The hybrid quantum-well/nanocrystal structures were excited at 266 nm by 200-fs pulses of the frequency-tripled output of an amplified Ti:sapphire laser. The emission from either the quantum well or the nanocrystals was selected using a monochromator. The dynamics measured for quantum-well/nanocrystal hybrid structures were compared with those in isolated quantum wells and isolated nanocrystal monolayers assembled on glass substrates. All measurements were performed at room temperature.

The interactions between the quantum well and the nanocrystal monolayer can be described in terms of a resonant Förster-type energy transfer¹¹. The energy-transfer rate per quantum-well carrier is strongly dependent on whether electrons and holes are free or bound by Coulomb interactions into excitons. In the case of excitons, the energy-transfer rate is independent of the density of quantum-well excitations (n_{eh}), whereas in the free-carrier case, the energy-transfer rate is proportional to n_{eh} (see Supplementary Information). To distinguish experimentally between these two cases, we measured the excitation-density dependence of photoluminescence in an isolated quantum well at $t = 0$ ps (Fig. 2). We observe that at low pump powers this dependence is quadratic, and it saturates at high excitation densities. The quadratic growth of photoluminescence is characteristic of free-carrier bimolecular recombination, indicating that electron–hole interactions in our quantum-well samples are not sufficiently strong to produce bound exciton states at room temperature. We also monitor the photoluminescence dynamics in the quantum well (Fig. 2b) and observe that the photoluminescence decay is exponential and is characterized by a time constant range of 0.6–1 ns that is independent of pump power. This result indicates that the decay of photoexcited carriers is dominated not by radiative recombination (characterized by the density-dependent time constant $\tau \propto 1/n_{\text{eh}}$) but by trapping at defects, as is typically observed for InGa_N quantum wells at room temperature.

After establishing that quantum-well excitations are unbound electron–hole pairs, we can analyse the energy-transfer rate, Γ_{ET} , per quantum-well carrier using the following expression (see

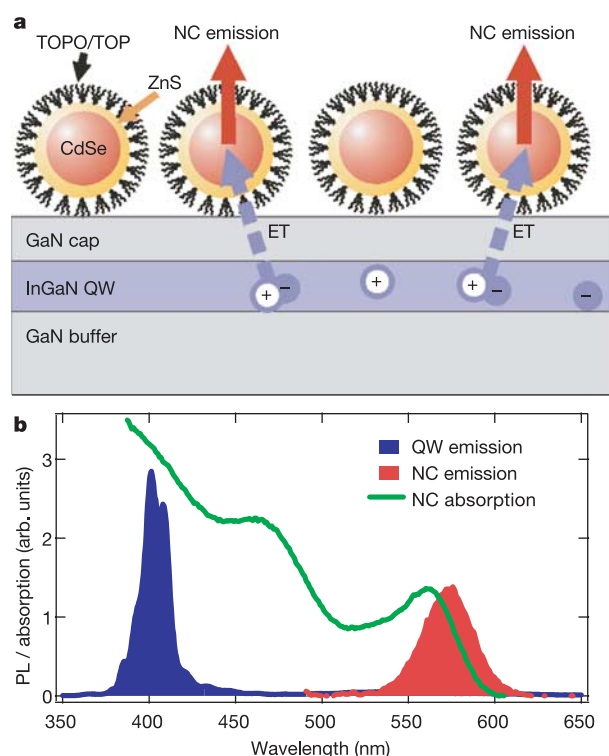


Figure 1 Schematic and optical properties of the hybrid quantum-well/nanocrystal structure. **a**, The structure consists of an InGa_N/GaN quantum-well heterostructure with a monolayer of TOPO/TOP-capped CdSe/ZnS core/shell nanocrystals on top of it. Electron–hole pairs generated in the quantum well can experience non-radiative resonant transfer into nanocrystals by Förster-type, dipole–dipole interactions. The nanocrystals excited by energy transfer produce emission with a wavelength determined by the nanocrystal size. **b**, The emission of the quantum well (blue) spectrally overlaps with the absorption of the nanocrystals (green). For CdSe nanocrystals with 1.9 nm radius, the emission wavelength is around 575 nm (red). PL is photoluminescence.

Supplementary Information):

$$\Gamma_{\text{ET}} = \frac{8\pi^2}{3\epsilon^2} |\mu_{\text{NC}}|^2 |\mu_{\text{QW}}|^2 n_{\text{NC}} n_{\text{eh}} N_{\text{NC}}(\hbar\omega_{\text{QW}}) \times \frac{1}{d^4} \frac{\hbar^2}{2Mk_{\text{B}}T} \int_0^\infty \kappa^3 \exp\left(-2\kappa - \frac{\hbar^2 \kappa^2}{2Mk_{\text{B}}Td^2}\right) d\kappa \quad (1)$$

in which ϵ is the dielectric constant, μ_{NC} and μ_{QW} are the transition dipole moments for the nanocrystal and the quantum well, respectively, n_{NC} is the surface density of nanocrystals, $N_{\text{NC}}(\hbar\omega_{\text{QW}})$ is the nanocrystal density of states at the quantum-well emission energy $\hbar\omega_{\text{QW}}$, d is the separation between the centres of the quantum well and the nanocrystal monolayer, M is the sum of electron and hole masses in the quantum well, T is the temperature, κ is the in-plane centre-of-mass momentum of the electron-hole pair in the quantum well in units of $\hbar/d$, and k_{B} is the Boltzmann constant. Assuming that the length of nanocrystal-passivating molecules is 1.1 nm, we find that d is 8.1 and 5.1 nm for capped and uncapped quantum wells, respectively, which further results in transfer rates of 1.05 ns^{-1} (capped quantum well) and 5.8 ns^{-1} (uncapped quantum well) at a quantum-well carrier density of $1.8 \times 10^{13} \text{ cm}^{-2}$. These estimated energy-transfer rates are sufficiently high to compete with carrier-decay rates ($\sim 1 \text{ ns}^{-1}$) measured experimentally for our quantum-well samples.

To measure directly the quantum-well-to-nanocrystal (QW-NC) energy-transfer rates, we perform comparative, time-resolved

photoluminescence studies for hybrid quantum-well/nanocrystal structures and isolated quantum wells. We observe that the presence of the nanocrystal layer adjacent to the quantum well significantly alters the quantum-well photoluminescence dynamics (Fig. 3a). Namely, the quantum-well photoluminescence decay becomes faster in the presence of nanocrystals, indicating an additional relaxation channel for quantum-well excitations, which is most likely due to QW-NC energy transfer (see the analysis below). This nanocrystal-induced change in quantum-well dynamics becomes more pronounced with increasing carrier density (compare traces shown by solid and dashed lines in Fig. 3a). To quantify this increase, in Fig. 3b we plot the additional initial decay rate $\Delta\Gamma = \Gamma_{\text{QW with NC}} - \Gamma_{\text{QW without NC}}$ as a function of n_{eh} for structures based on uncapped and capped quantum wells. We observe that in both cases the $\Delta\Gamma$ growth is linear with n_{eh} , but absolute values of $\Delta\Gamma$ are approximately 4.4 times greater for the uncapped quantum wells compared with quantum wells with a top barrier. Both of these observations are consistent with the fact that the additional decay rate $\Delta\Gamma$ is due to QW-NC energy transfer. Förster modelling (equation (1)) predicts that for the free-carrier case, the energy-transfer rate should increase linearly with n_{eh} , which is exactly the dependence observed experimentally. Furthermore, the increase in the transfer rate in the case of the uncapped quantum well is consistent with its strong dependence on the energy-transfer distance ($\Gamma_{\text{ET}} \sim d^{-4}$). From the geometrical parameters of our system, we estimate that the d dependence should result in a 5.5

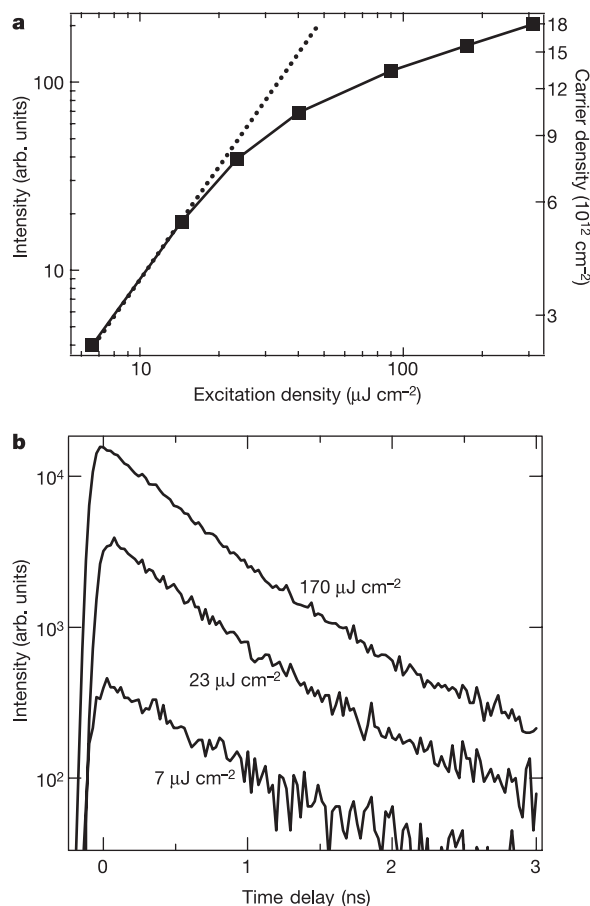


Figure 2 Pump- and time-dependent emission from an isolated quantum well. **a**, The photoluminescence intensity of the quantum well (squares) at zero time delay is plotted as a function of pump fluence. The dotted line is a fit to the quadratic growth at low pump fluences. **b**, Photoluminescence dynamics of the isolated quantum well measured at different pump fluences.

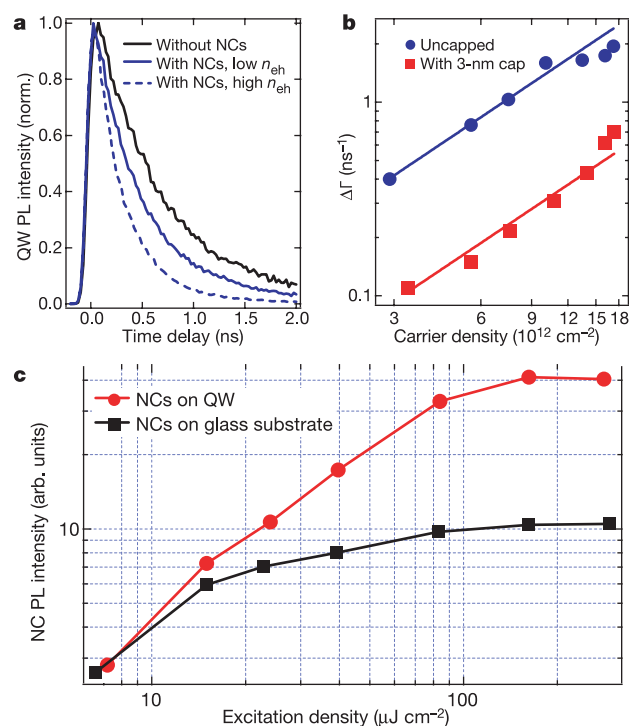


Figure 3 Experimental observations of QW-NC energy transfer. **a**, Normalized photoluminescence dynamics of the isolated quantum well (black solid line) at $n_{\text{eh}} = 3 \times 10^{12} \text{ cm}^{-2}$ in comparison with quantum-well photoluminescence dynamics measured for the quantum-well/nanocrystal structure at $n_{\text{eh}} = 3 \times 10^{12} \text{ cm}^{-2}$ (blue solid line) and $n_{\text{eh}} = 10 \times 10^{12} \text{ cm}^{-2}$ (blue dashed line). **b**, The difference between the initial photoluminescence decay rates measured for the isolated quantum well and the quantum-well/nanocrystal structure ($\Delta\Gamma = \Gamma_{\text{QW with NC}} - \Gamma_{\text{QW without NC}}$) versus quantum-well carrier density for samples based on the capped and the uncapped quantum well. **c**, Time-integrated nanocrystal photoluminescence intensity versus pump fluence for the nanocrystal monolayer assembled on a glass substrate and on top of a capped quantum well.

increased energy-transfer rate, which agrees well with the factor of 4.4 observed experimentally.

Further evidence for efficient QW–NC energy-transfer is provided by the analysis of the photoluminescence from the nanocrystal layer. The energy-transfer-induced outflow of carriers from the quantum well should result in a corresponding increase in the emission of the nanocrystals. In our experimental configuration, however, optical pumping directly generates carriers not only in the quantum well but also in the nanocrystals. Therefore, to extract the energy-transfer-induced increase in the nanocrystal photoluminescence, we perform a side-by-side comparison of photoluminescence data for hybrid quantum-well/nanocrystal structures and a nanocrystal Langmuir–Blodgett monolayer assembled on a glass slide. One such set of data plotted as temporally integrated nanocrystal photoluminescence intensity versus pump fluence is displayed in Fig. 3c. To account for the difference in the nanocrystal packing densities for

Langmuir–Blodgett films assembled on the quantum well and the glass slide, we introduce a constant scaling factor, which allows us to match photoluminescence intensities detected from quantum-well/nanocrystal and glass/nanocrystal samples at low pump powers, for which energy transfer from the quantum well is negligible. The data indicate that at low pump fluences, both types of samples show a similar photoluminescence pump dependence. However, two traces show distinctly different behaviour at higher pump fluences, for which energy transfer starts to play a significant role (Fig. 3b). Whereas emission from the isolated nanocrystal layer saturates at $\sim 20 \mu\text{J cm}^{-2}$, the nanocrystal photoluminescence in the hybrid structure shows a steady growth until $\sim 80 \mu\text{J cm}^{-2}$. As a result of this delayed saturation, the maximum nanocrystal photoluminescence intensity achievable with the quantum-well/nanocrystal structure is four times greater than the photoluminescence for the nanocrystal monolayer on the glass slide. All of these results indicate a strong additional carrier inflow into nanocrystals as a result of energy transfer from the quantum well.

Figure 4a displays the schematics of energy transfer along with other relaxation processes in the hybrid quantum-well/nanocrystal structures studied in this work. Following photoexcitation, carrier thermalization, and cooling, the thermal distribution of free electrons and holes is established in the quantum well. Quantum-well carriers can decay either radiatively (time constant τ_r) or non-radiatively (τ_{nr}), or experience energy transfer (τ_{ET}) into a nanocrystal. Carriers generated in the nanocrystal by resonant QW–NC energy transfer have significant excess energies as measured with respect to the nanocrystal bandgap. Extremely fast intraband relaxation in nanocrystals (subpicosecond time scales)^{12,13} rapidly removes carriers from resonance with the quantum-well transition and prevents backtransfer. In well-passivated nanocrystals, relaxed electron–hole pairs recombine primarily radiatively with a time constant of ~ 20 ns, emitting a photon with an energy that is determined by the nanocrystal size.

The efficiency of non-radiative QW–NC energy-transfer (η_{ET}) can be estimated from the expression: $\eta_{ET} = \tau_r (\tau_{ET} + \tau_r)^{-1}$, in which $\tau_r = (1/\tau_{nr} + 1/\tau_r)^{-1}$ is the relaxation time of quantum-well excitations due to both radiative and non-radiative process. Our experimental results for the uncapped sample indicate that $\tau_r \approx 0.6$ ns and $\tau_{ET} \approx 0.5$ ns (for $n_{eh} = 1.8 \times 10^{13} \text{ cm}^{-2}$), which yields η_{ET} as high as 55%. We believe that nearly 100% efficiencies can be achieved by improving the quality of the quantum wells (to reduce non-radiative losses) and/or by optimizing the geometry of the nanocrystal/quantum-well structure (by using, for example, shorter nanocrystal surface-passivation molecules).

It is interesting that despite the additional step in the energy-transfer process, the photoluminescence quantum yield of the hybrid quantum-well/nanocrystal device ($QY_{QW/NC}$) can be greater than the original quantum yield of the quantum well. $QY_{QW/NC}$ can be estimated from the expression $QY_{QW/NC} = QY_{NC} (1 + \tau_{ET}/\tau_r)^{-1}$. This expression indicates that if $\tau_{ET} \ll \tau_r$ the quantum efficiency of the hybrid structure approaches that of nanocrystals. This conclusion further means that even the use of InGaN quantum wells with poor room-temperature quantum yields can produce highly efficient hybrid devices.

It is illustrative to compare the efficiency of energy transfer measured here with that expected for radiative energy transfer (η_{RET}). The latter process is used in the traditional colour-conversion scheme, and is based on the emission of a photon from a quantum well followed by absorption/re-emission steps in the phosphor material (nanocrystals in our case). For a close-packed nanocrystal monolayer, η_{RET} can be estimated from the ratio of the nanocrystal absorption cross-section¹⁴ to its geometrical cross-section, which yields $\eta_{RET} < 0.3\%$. This value is at least two orders of magnitude smaller than the efficiencies measured experimentally, indicating that the use of non-radiative energy transfer

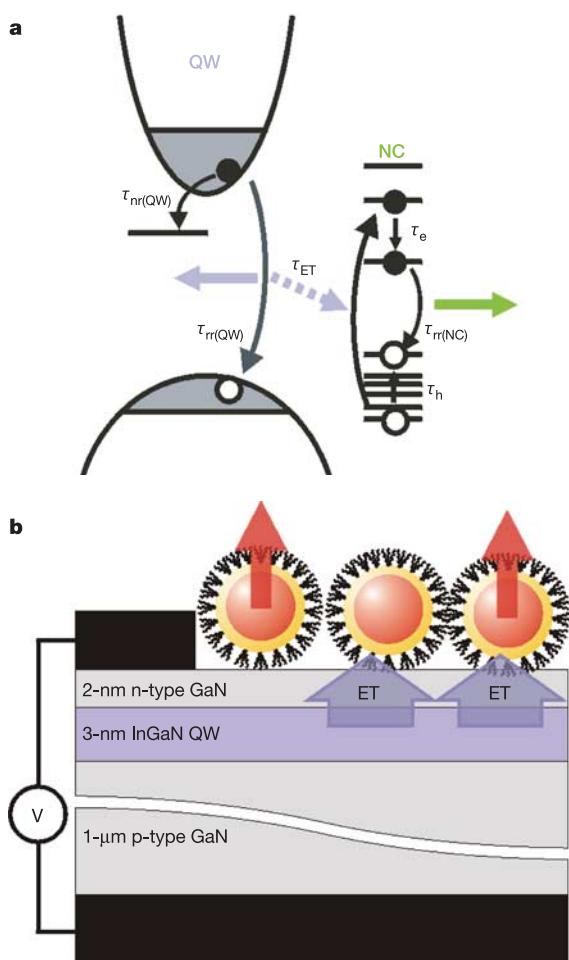


Figure 4 Carrier relaxation and energy-transfer processes in the hybrid quantum-well/nanocrystal structure and a schematic of an electrically driven energy-transfer device.

a, The QW–NC energy transfer competes with radiative and non-radiative decay processes in the quantum well. High-energy excitations created in the nanocrystals through energy transfer rapidly relax (time constants τ_e and τ_h) to the nanocrystal band edge, which prevents backtransfer. Subscript e is electron and h is hole. **b**, An electrically powered hybrid quantum-well/nanocrystal device that can be used to realize the ‘energy-transfer colour-converter’ in the regime of electrical injection. It depicts an InGaN quantum well sandwiched between bottom p-type and thin, top n-type GaN barriers with attached metal contacts. The top contact only partially covers the quantum well and leaves open space for assembling the nanocrystals.

can significantly improve the performance of colour-conversion devices.

In addition to applications as efficient colour converters, nanocrystals have been considered promising building blocks for colour-selectable optical-gain media in lasing applications⁴. One complication associated with lasing applications of nanocrystals is the requirement for extremely fast pumping that competes with non-radiative Auger recombination, leading to very short (picosecond) optical gain lifetimes¹⁵. So far, optical amplification and lasing in nanocrystals has been achieved using optical excitation with short laser pulses. Our estimations show that the 'energy-transfer pumping' scheme studied here provides carrier inflow that can in principle compete with non-radiative losses induced by Auger recombination. The energy-transfer rate of $\sim 2 \text{ ns}^{-1}$ measured for the uncapped quantum-well sample for $n_{\text{ch}} = 1.8 \times 10^{13} \text{ cm}^{-2}$ results in a QW-NC carrier flux of $\sim 3.6 \times 10^{22} \text{ cm}^{-2} \text{ s}^{-1}$. For the nanocrystals of 1.9 nm radius studied here, the Auger recombination time is $\sim 50 \text{ ps}$, which corresponds to a non-radiative carrier loss of $4 \times 10^{22} \text{ cm}^{-2} \text{ s}^{-1}$ for a close-packed monolayer. The latter value is comparable to the carrier inflow rate provided by energy transfer from the quantum well, indicating the feasibility of lasing in the energy-transfer pumping regime.

Although in this report we have studied optically pumped devices, it should be possible to realize the energy-transfer pumping scheme in the regime of electrical injection by combining nanocrystals with an electrically driven InGaN quantum well. The design of the quantum-well emitter in the 'energy-transfer colour-converter' (Fig. 4b) can be similar to that used in conventional InGaN light-emitting diodes, in which the quantum well is sandwiched between n- and p-doped GaN barriers with attached metal contacts¹⁶. Our preliminary studies indicate that we can fabricate relatively high mobility, thin (2–3 nm), n-doped GaN layers that can be used as top quantum-well barriers (adjacent to nanocrystals) in electrically powered devices. The direct comparison of photoluminescence dynamics in nanocrystals assembled on glass slides and n-doped GaN layers (up to $2 \times 10^{19} \text{ cm}^{-3}$ doping level) do not show any noticeable quenching of nanocrystal emission in the presence of a proximal, doped semiconductor. Furthermore, the doping of the barriers is not expected to induce additional carrier losses in the quantum well^{17,18}. All of these considerations strongly indicate the feasibility of high-efficiency, electrically driven, hybrid nanocrystal/quantum-well devices. □

Received 9 February; accepted 19 April 2004; doi:10.1038/nature02571.

1. Alivisatos, A. P. Semiconductor clusters, nanocrystals, and quantum dots. *Science* **271**, 933–937 (1996).
2. Murray, C. B., Norris, D. J. & Bawendi, M. G. Synthesis and characterization of nearly monodisperse CdE (E = S, Se, Te) semiconductor nanocrystallites. *J. Am. Chem. Soc.* **115**, 8706–8715 (1993).
3. Bruchez, M., Moronne, M., Gin, P., Weiss, S. & Alivisatos, A. P. Semiconductor nanocrystals as fluorescent biological labels. *Science* **281**, 2013–2016 (1998).
4. Klimov, V. I. *et al.* Optical gain and stimulated emission in nanocrystal quantum dots. *Science* **290**, 314–317 (2000).
5. Next-Generation Lighting Initiative <http://lighting.sandia.gov/Xlightinginit.htm#snlgl>
6. Colvin, V., Schlamp, M. & Alivisatos, A. Light-emitting diodes made from cadmium selenide nanocrystals and a semiconducting polymer. *Nature* **370**, 354–357 (1994).
7. Schlamp, M. C., Peng, X. G. & Alivisatos, A. P. Improved efficiencies in light emitting diodes made with CdSe(CdS) core/shell type nanocrystals and a semiconducting polymer. *J. Appl. Phys.* **82**, 5837–5842 (1997).
8. Coe, S., Woo, W.-K., Bawendi, M. & Bulovic, V. Electroluminescence from single monolayers of nanocrystals in molecular organic devices. *Nature* **420**, 800–803 (2002).
9. Dabbousi, B. O. *et al.* (CdSe)ZnS core-shell quantum dots: synthesis and characterization of a size series of highly luminescent nanocrystallites. *J. Phys. Chem.* **101**, 9463–9475 (1997).
10. Koleske, D. D. *et al.* Improved brightness of 380 nm GaN light emitting diodes through intentional delay of the nucleation island coalescence. *Appl. Phys. Lett.* **81**, 1940–1942 (2002).
11. Basko, D., La Rocca, G. C., Bassani, F. & Agranovich, V. M. Förster energy transfer from a semiconductor quantum well to an organic material overlayer. *Eur. Phys. J. B* **8**, 353–362 (1999).
12. Klimov, V. I. & McBranch, D. W. Femtosecond IP-to-1S electron relaxation in strongly confined semiconductor nanocrystals. *Phys. Rev. Lett.* **80**, 4028–4031 (1998).
13. Xu, S., Mikhailovsky, A. A., Hollingsworth, J. A. & Klimov, V. I. Hole intraband relaxation in strongly confined quantum dots: Revisiting the 'phonon bottleneck' problem. *Phys. Rev. B* **65**, 045319 (2002).
14. Klimov, V. I. Optical nonlinearities and ultrafast carrier dynamics in semiconductor nanocrystals. *J. Phys. Chem. B* **104**, 6112–6123 (2000).

15. Klimov, V. I., Mikhailovsky, A. A., McBranch, D. W., Leatherdale, C. A. & Bawendi, B. G. Quantization of multiparticle Auger rates in semiconductor quantum dots. *Science* **287**, 1011–1013 (2000).
16. Nakamura, S. & Fasol, G. *The Blue Laser Diode: GaN Based Light Emitters and Lasers* (Springer, Berlin Heidelberg, Germany, 1997).
17. Wu, L. W. *et al.* Influence of Si-doping on the characteristics of InGaN-GaN multiple quantum-well blue light emitting diodes. *IEEE J. Quantum Electron.* **38**, 446–450 (2002).
18. Bidnyk, S. *et al.* High-temperature stimulated emission in optically pumped InGaN/GaN multiquantum wells. *Appl. Phys. Lett.* **72**, 1623–1625 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by Los Alamos LDRD Funds and the Office of Basic Energy Sciences, Office of Science, US Department of Energy. Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the Department of Energy.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to V.I.K. (klimov@lanl.gov) or M.A. (achermann@lanl.gov).

High levels of atmospheric carbon dioxide necessary for the termination of global glaciation

Raymond T. Pierrehumbert

Department of the Geophysical Sciences, The University of Chicago, 5734 South Ellis Avenue, Chicago, Illinois 60637, USA

The possibility that the Earth suffered episodes of global glaciation as recently as the Neoproterozoic period, between about 900 and 543 million years ago, has been widely discussed^{1–3}. Termination of such 'hard snowball Earth' climate states has been proposed to proceed from accumulation of carbon dioxide in the atmosphere⁴. Many salient aspects of the snowball scenario depend critically on the threshold of atmospheric carbon dioxide concentrations needed to trigger deglaciation^{2,5}. Here I present simulations with a general circulation model, using elevated carbon dioxide levels to estimate this deglaciation threshold. The model simulates several phenomena that are expected to be significant in a 'snowball Earth' scenario, but which have not been considered in previous studies with less sophisticated models, such as a reduction of vertical temperature gradients in winter, a reduction in summer tropopause height, the effect of snow cover and a reduction in cloud greenhouse effects. In my simulations, the system remains far short of deglaciation even at atmospheric carbon dioxide concentrations of 550 times the present levels (0.2 bar of CO₂). I find that at much higher carbon dioxide levels, deglaciation is unlikely unless unknown feedback cycles that are not captured in the model come into effect.

Whereas the problem of initiation of a 'hard snowball' climate state has received detailed attention^{6–8}, most current thinking about deglaciation is based on highly idealized energy balance model (EBM) calculations⁹, which offer an accurate treatment of clear-sky radiation but neglect the seasonal cycle and fix cloud radiative forcing at its present value. Although 0.12 bar of CO₂ is often quoted as a representative deglaciation threshold, a closer reading of the work yields a threshold of 0.29 bar based on Neoproterozoic insolation (Fig. 2 of ref. 9). The same model but with slightly different choices of parameters¹⁰ achieves deglaciation at only 0.16 bar. Because of the weak logarithmic dependence of radiative forcing on CO₂, the EBM results consistently imply that the system should be at least close to deglaciation at 0.2 bar of CO₂.

Four climatic characteristics crucially affect deglaciation, but are exceedingly difficult to estimate in simplified models. They are temperature lapse rate, cloud effects, dynamical heat transport and snow cover. Lapse rate, governed by a complex interplay of dynamics and convection, is crucial to the greenhouse effect, which can operate only insofar as the air aloft is significantly colder than the ground. Clouds are important because a high cloud over ice or snow reflects little more sunlight than the underlying surface, but provides a powerful warming effect through trapping of infrared radiation. EBM studies confirm the sensitivity of the deglaciation threshold to cloud assumptions¹¹. Horizontal heat transport by atmospheric motion determines the extent to which the warm areas, which are the first to deglaciate, must give up some of their energy to the colder parts of the planet. Snow cover, determined by long-range transport of moisture in the atmosphere, affects the surface albedo, because snow is more reflective than sea ice¹².

The following results are based on general circulation model (GCM) simulations as described in the Methods section. The climate of the hard-snowball Earth is governed by the low thermal inertia of the globally solid surface, which has the consequence that the temperature responds primarily to the instantaneous solar radiation. The summer hemisphere becomes nearly isothermal, whereas the weakly illuminated winter hemisphere becomes extraordinarily cold, resulting in an extreme seasonal cycle resembling that of Mars^{13,14}. Consider the January zonal mean near-surface air temperature for 100 p.p.m. CO₂ (Fig. 1). The south (summer) polar temperature is 228 K, only about 17 K cooler than the tropics, while the north (winter) polar temperature falls to 163 K. In July, the pattern is much the same, except reflected about the Equator, yielding a 64 K high-latitude seasonal cycle. During the equinoxes (not shown) the temperature is nearly symmetric about the Equator, with equatorial temperatures somewhat cooler than the summer subtropics, and polar temperatures somewhat warmer than those seen in Fig. 1 at the winter pole. Here I present results only for the solstice conditions, as a nearly identical discussion applies to the equinox.

CO₂ increase yields little warming in the winter hemisphere, and the summer temperature remains well short of the freezing point even at 0.2 bar of CO₂ (Fig. 1). Why is the warming so weak? Fig. 2 shows the diagnosed clear-sky greenhouse effect, defined as $G = \sigma T_s^4 - \text{OLR}_{\text{clear}}$, where T_s is the surface temperature and $\text{OLR}_{\text{clear}}$ is the clear-sky outgoing longwave radiation. It is actually negative in

the winter extratropics, and grows only to modest values in the winter tropics. The reason for this behaviour is to be found in the vertical profile of temperature (Fig. 3). In the winter hemisphere, the atmosphere is nearly isothermal, because, in the absence of convection due to solar heating or a warm ocean, the atmosphere relaxes to radiative equilibrium, with a temperature inversion at the surface. Without colder air aloft, the winter hemisphere acts somewhat like the present-day stratosphere, which experiences a radiative cooling tendency in response to an increase of the CO₂ concentration. In the summer hemisphere, the low tropopause limits the vertical temperature contrast, yielding a weak greenhouse effect compounded by the virtual lack of water vapour feedback at such cold temperatures. As a result, at 0.2 bar of CO₂ the greenhouse trapping never attains even the 100 W m⁻² typical of the present low-CO₂ climate.

The cloud greenhouse effect (Fig. 2) is also weak, for two robust reasons: (1) the cloud greenhouse effect arises from high clouds, but even at saturation there is little water in the cold upper summer atmosphere, or at any level in the winter hemisphere; (2) the weak summer meridional temperature gradient cannot support the baroclinic eddies that lead to mid-latitude storm-track clouds in the modern climate. Significant cloud cover is limited primarily to the upward branch of the Hadley circulation, around 20° in the summer hemisphere. The cloud greenhouse effect barely exceeds 10 W m⁻², as compared to approximately 90 W m⁻² in the modern climate.

Dynamical heat transport due to the Hadley cell, and to a storm track at 30° in the winter hemisphere, draws energy out of the summer hemisphere. In January, the effect is equivalent to 45 W m⁻² of radiative cooling at 25°S in the 0.2 bar case.

Snow cover is a potent impediment to deglaciation. The tropical ice from about 10° N to 10° S is an ablation zone in all cases, losing water by sublimation. This water is accumulated as snow on the rest of the planet. The net snow accumulation is only about 0.5 cm liquid water equivalent per year, but over a mere decade the snow layer becomes deep enough to increase the albedo over most of the planet's surface. The interior of the tropical supercontinent remains devoid of snow, and warms the tropics through a reduction of the tropical surface albedo. The tropics are probably spuriously warm in this simulation. The bare tropical sea-ice is an artefact of the neglect of ice dynamics, as the extratropical snow and ice accumulation would eventually lead to sea-glacier flow, replacing the tropical sea ice with brighter glacial ice with an albedo of about 0.6 (refs 12, 15).

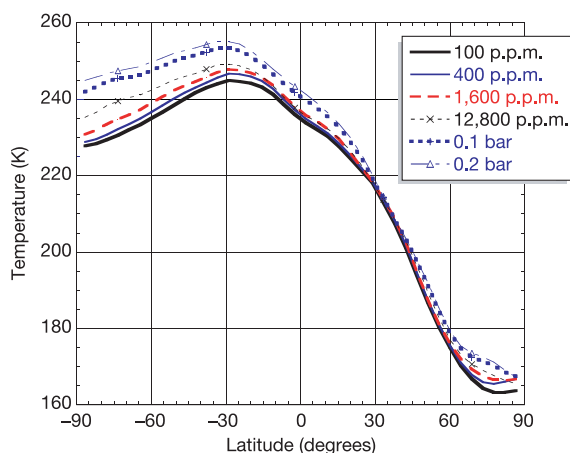


Figure 1 January zonal-mean air temperature at the lowest model level, for various concentrations of atmospheric CO₂. Only sea-ice grid points are used in computing the mean, so as to focus on the temperature most relevant for determining deglaciation.

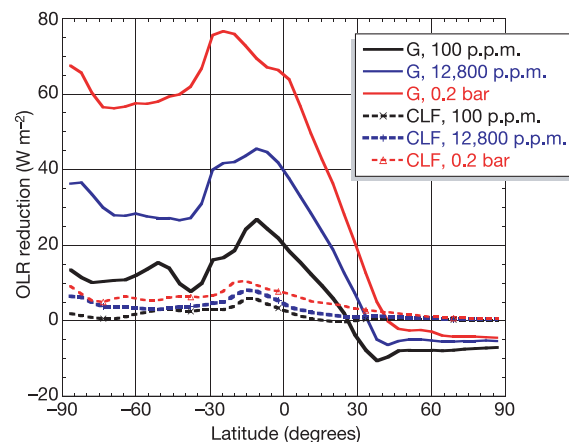


Figure 2 January zonal-mean clear-sky greenhouse trapping (solid lines) and cloud longwave forcing (CLF, dashed lines), for various CO₂ concentrations. The cloud longwave forcing is defined as the reduction in OLR caused by cloud effects, beyond the reduction caused by the clear-sky greenhouse effect.

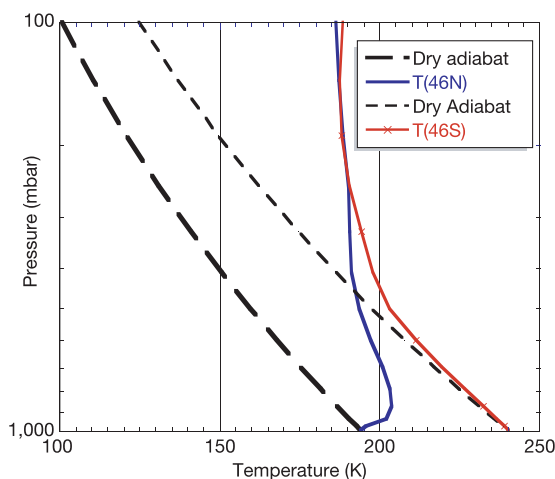


Figure 3 Typical January vertical temperature profiles for the 100 p.p.m. CO_2 case. For reference, the dry adiabat (virtually identical to the moist adiabat at these temperatures) is shown. Convection resets the temperature to the adiabat, so by comparing the actual curves with the adiabat, we can identify the layer in which convection causes the familiar sharp drop of temperature with height. $T(46\text{N})$ and $T(46\text{S})$ represent respectively the temperature profiles at latitudes 46°N and 46°S .

Land glacier dynamics, neglected in this simulation, might cause the continent to become glaciated as well, leading to further cooling¹⁶.

The aggregate of all the effects discussed above leaves the maximum air temperature 18 K short of the melting point at 0.2 bar; the ice surface temperature is nearly 10 K colder, owing to strong night-time cooling, and the equatorial deep ice temperature (which tracks the annual average) is colder still, resulting in the system being at least 30 K short of deglaciation. The present GCM cannot be reliably used above 0.2 bar, but if no new physical process enters to increase climate sensitivity, a logarithmic extrapolation suggests that each further doubling would warm the climate by about 2 K. At this rate, even 3.2 bar of CO_2 would be insufficient to trigger deglaciation. Assuming present outgassing rates¹⁷ and assuming that two-thirds of outgassed CO_2 goes into the ocean⁵, it would take 28 million years for 0.2 bar to accumulate in the atmosphere; accumulation of 1 or 2 bar would certainly not be impossible, but it would strain the limits of possibility.

It is far from certain that a hard-snowball state ever actually occurred on Earth, but the difficulty of deglaciation does not in itself rule out the possibility. There are several poorly understood physical feedbacks that may yet permit deglaciation. Alternative cloud parameterizations may retain more cloud water in the atmosphere. Also, above 0.2 bar, formation of CO_2 clouds becomes important and these can have a pronounced warming effect¹⁸. Greenhouse gases other than CO_2 may be important, or surface albedo may be reduced by admixture of dust, or by formation of leads in sea-glacier fracture zones or perhaps in thin-ice regions maintained by hydrothermal plumes¹⁹. Even the suppressed convection that leads to weak vertical temperature contrast depends to some extent on surface flux parameterizations, the representation of the diurnal cycle and the convection scheme; future advances in the understanding of cold-climate convection may yield different behavior. It is in this veiled area of physics that the prospect for recovery from a hard snowball lies hidden. $\square$

Methods

The simulations were carried out using the FOAM 1.5 GCM^{8,20}. Once sea ice has built up to sufficient thickness, ocean dynamics has little effect. Hence, the simulations were carried out with a mixed-layer ocean without imposed horizontal oceanic heat transport. Run in this mode, FOAM is essentially a portable Beowulf-oriented re-implementation of

CCM3²¹, run at R15 horizontal resolution ($4.5^\circ \times 7.5^\circ$) with 18 levels. The solar luminosity was set at 94% of its present value, and the palaeogeography consists of an idealized equatorial supercontinent as in refs 8, 14. In the hard snowball, continental configuration affects climate almost exclusively through surface albedo, and even then only to the extent that land remains snow-free. The assumed configuration is highly favourable to deglaciation, as it puts dark bare land in the tropics where it can optimally receive insolation and transfer energy to the rest of the tropics. Orbital parameters were left at the Earth's present values. Surface albedo is specified in two spectral bands, with a broadband average value of 0.75 for snow (somewhat less than the new-snow value in ref. 12) and 0.5 for bare sea ice¹².

The model was first run at 100 p.p.m. CO_2 until the ocean became globally covered by sea ice of thickness of at least 5 m. Then, a sequence of 20-year simulations was carried out, with CO_2 concentration set at 400 p.p.m., 1,600 p.p.m., 1,2800 p.p.m., 10% (0.1 bar) and 20% (0.2 bar). Given the low thermal inertia of the ice/land surface, 20 years was found to be adequate for the climate to come into equilibrium with each CO_2 concentration. Results are taken from the last 10 years of each run. The only quantities that fail to reach equilibrium are the ice thickness and snow depth, but in neither case does the disequilibrium affect the climate. Snow continues to accumulate at up to 0.5 cm per year of liquid water equivalent, but as surface albedo saturates at a snow depth of only 0.5 cm over ice, the surface albedo has ample time to equilibrate in regions of net accumulation. To ensure that the ice/snow model retains adequate vertical resolution to treat the diurnal and seasonal cycle, ice thickness is clamped at a maximum of 20 m, and snow thickness at a maximum of 1 m (liquid water equivalent). This value is reached quickly, and given the low thermal diffusivity of ice, is sufficient to almost completely insulate the atmosphere from the heat content of the ocean. If allowed to grow, ice thickness would continue to increase at a rate of 20–50 cm yr^{-1} , further reducing the already small leakage of heat through the ice; the leakage through 20 m of ice causes the climate to be very slightly warmer than it would be if ice were allowed to reach its full thickness.

Although 1,600 p.p.m. is well within the range for which the CCM3 radiation code in FOAM is validated, there is little published basis on which to estimate the errors for CO_2 levels as high as 0.1 bar. To address this issue, we used a radiation model valid at high CO_2 (ref. 22) to recompute the outgoing longwave radiation (OLR) corresponding to a representative sampling of the temperature and humidity profiles produced by the GCM. The accurate OLR was then compared with the OLR produced by the GCM's internal radiation code. A typical comparison is shown in Supplementary Information. At 0.2 bar, the GCM radiation code over-estimates the correct OLR by at most 3.8 W m^{-2} in the warmest months, with a mid-latitude annual mean error of 1.9 W m^{-2} .

Received 12 November 2003; accepted 7 May 2004; doi:10.1038/nature02640.

- Hoffman, P. F., Kaufman, A. J., Halverson, G. P. & Schrag, D. P. A Neoproterozoic snowball earth. *Science* **281**, 1342–1346 (1998).
- Hoffman, P. F. & Schrag, D. P. The snowball Earth hypothesis: testing the limits of global change. *Terra Nova* **14**, 129–155 (2002).
- Kennedy, M. J., Christie-Blick, N. & Sohl, L. E. Are Proterozoic cap carbonates and isotopic excursions a record of gas hydrate destabilization following Earth's coldest intervals? *Geology* **29**(5), 443–446 (2001).
- Kirschvink, J. L. In *The Proterozoic Biosphere; a Multidisciplinary Study* (eds Schopf, J. W. & Klein, C.) 51–52 (Cambridge Univ. Press, Cambridge, UK, 1992).
- Higgins, J. A. & Schrag, D. P. Aftermath of a snowball Earth. *Geochim. Geophys. Geosyst.* **4**(3), doi:10.1029/2002GC000403 (2003).
- Chandler, M. A. & Sohl, L. E. Climate forcings and the initiation of low-latitude ice sheets during the Neoproterozoic Varanger glacial interval. *J. Geophys. Res.* **D 105**, 20737–20756 (2000).
- Hyde, W. T., Crowley, T. J., Baum, S. K. & Peltier, W. R. Neoproterozoic "snowball Earth" simulations with a coupled climate/ice-sheet model. *Nature* **405**, 425–429 (2000).
- Poulsen, C., Pierrehumbert, R. T. & Jacob, R. Impact of ocean dynamics on the simulation of the Neoproterozoic "Snowball Earth". *Geophys. Res. Lett.* **28**(8), 1575–1578 (2001).
- Caldeira, K. & Kasting, J. F. Susceptibility of the early Earth to irreversible glaciation caused by carbon dioxide clouds. *Nature* **359**, 226–228 (1992).
- Tajika, E. Faint young Sun and the carbon cycle: implication for the Proterozoic global glaciations. *Earth Planet. Sci. Lett.* **214**, 443–453 (2003).
- Pierrehumbert, R. T. The hydrologic cycle in deep-time climate problems. *Nature* **419**, 191–198 (2002).
- Warren, S. G., Brandt, R. E., Grenfell, T. C. & McKay, C. P. Snowball Earth: ice thickness on the tropical ocean. *J. Geophys. Res.* **C 107**, doi:10.1029/2001JC001123 (2002).
- Walker, J. C. G. *Earth System Processes*, Abstr 110–111 (Geological Society, London, 2001).
- Jenkins, G. S. GCM greenhouse and high-obliquity solutions for early Proterozoic glaciation and middle Proterozoic warmth. *J. Geophys. Res.* **D 108**, doi:10.1029/2001JD001582 (2003).
- Goodman, J. C. & Pierrehumbert, R. T. Glacial flow of floating marine ice in Snowball Earth. *J. Geophys. Res.* **C 108**, doi:10.1029/2002JC001471 (2003).
- Donnadieu, Y., Fluteau, F., Ramstein, G., Ritz, C. & Besse, J. Is there a conflict between the Neoproterozoic glacial deposits and the snowball Earth interpretation: an improved understanding with numerical modeling. *Earth Planet. Sci. Lett.* **208**, 101–112 (2003).
- Zhang, Y. X. & Zindler, A. Distribution and evolution of carbon and nitrogen in Earth. *Earth Planet. Sci. Lett.* **117**, 331–345 (1993).
- Forget, F. & Pierrehumbert, R. T. Warning early Mars with carbon dioxide clouds that scatter infrared radiation. *Science* **278**, 1273–1276 (1997).
- Goodman, J. C., Collins, G. C., Marshall, J. & Pierrehumbert, R. T. Hydrothermal plume dynamics on Europa: Implications for chase formation. *J. Geophys. Res.* **E 109**, doi:10.1029/2003JE002073 (2004).
- Jacob, R. *Low Frequency Variability in a Simulated Atmosphere Ocean System*. Thesis, Univ. Wisconsin-Madison (1997).
- Kiehl, J. T. et al. The National Center for Atmospheric Research Community. Climate Model: CCM3. *J. Clim.* **11**, 1131–1149 (1998).

22. Kasting, J. F., Pollack, J. B. & Ackerman, T. P. Response of Earth's atmosphere to increases in solar flux and implications for loss of water from Venus. *Icarus* **57**, 335–355 (1984).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank P. Hoffman and S. Warren for discussions on a range of matters relating to the Neoproterozoic and to surface albedo in general; J.C.G. Walker for sharing additional thoughts concerning the Mars analogy; and LMD/Paris for providing a congenial environment in which to carry out this work. This work was funded by the National Science Foundation.

Competing interests statement The author declares that he has no competing financial interests.

Correspondence and requests for materials should be addressed to the author (rtp1@geosci.uchicago.edu).

Mesozoic origin for West Indian insectivores

Alfred L. Roca^{1*}, Gila Kahila Bar-Gal^{2*}, Eduardo Eizirik^{2,3}, Kristofer M. Helgen⁴, Roberto Maria⁵, Mark S. Springer⁶, Stephen J. O'Brien² & William J. Murphy¹

¹Laboratory of Genomic Diversity, Basic Research Program, SAIC-Frederick and
²Laboratory of Genomic Diversity, National Cancer Institute, Frederick, Maryland 21702, USA

³Centro de Biologia Genômica e Molecular, PUCRS, Porto Alegre, Brazil

⁴School of Earth and Environmental Sciences, University of Adelaide, Adelaide 5005, Australia

⁵Parque Zoológico Nacional, ZOODOM, Santo Domingo, Dominican Republic

⁶Department of Biology, University of California, Riverside, California 92521, USA

* These authors contributed equally to this work

The highly endangered solenodons, endemic to Cuba (*Solenodon cubanus*) and Hispaniola (*S. paradoxus*), comprise the only two surviving species of West Indian insectivores^{1,2}. Combined gene sequences (13.9 kilobases) from *S. paradoxus* established that solenodons diverged from other eulipotyphlan insectivores 76 million years ago in the Cretaceous period, which is consistent with vicariance, though also compatible with dispersal. A sequence of 1.6 kilobases of mitochondrial DNA from *S. cubanus* indicated a deep divergence of 25 million years versus the congeneric *S. paradoxus*, which is consistent with vicariant origins as tectonic forces separated Cuba and Hispaniola^{3,4}. Efforts to prevent extinction of the two surviving solenodon species would conserve an entire lineage as old or older than many mammalian orders.

Solenodons are small (1 kg) fossorial (burrowing) insectivores, and are among the few native non-flying mammals that survived human settlement of the islands of the West Indies^{1,2}. They inhabit the forests of Cuba and Hispaniola to elevations of 2,000 m, and shelter in caves, crevices, logs and extensive tunnel networks at a depth of >20 cm (refs 5 and 6). The dearth of Late Cretaceous or early Tertiary fossils from the West Indies has constrained resolution among alternative hypotheses regarding the origin of solenodons and their affinity to other mammals^{1,7}.

Some have suggested a close relationship to soricids (shrews) but not to talpids (moles)^{8,9}, or to soricids but not erinaceids (hedgehogs and gymnures)^{9–11}, and/or to fossil North American 'apternodontids' such as *Apternodus*, or geolabids such as *Centetodon*^{10,12–15}. A few authorities have suggested an affinity of solenodons to Afro-Malagasy tenrecs (both have zalamdodont molars)^{12,15,16} and a trans-Atlantic dispersal event was suggested to explain this apparent relationship^{12,16}. Recent molecular studies have placed the tenrecs firmly within Afrotheria, a superordinal mammalian group with

African origins^{17,18}, while placing shrews, moles and erinaceids in a distinct clade (Eulipotyphla) within Laurasiatheria, a superordinal mammalian group most probably of Northern Hemisphere origins¹⁸. For solenodons, only a few mtDNA sequences of *S. paradoxus* have been available for analyses; these have rejected a close affinity between solenodons and tenrecs¹⁷. One study has placed *Solenodon* as a sister group to soricids + talpids but not to erinaceids, although the bootstrap support for this placement (51%) was quite weak¹⁷; a second molecular analysis has positioned *Solenodon* as sister to a clade of rodents¹⁹.

To examine the origin of *Solenodon* and its relationship to other mammals, we sequenced portions of 16 nuclear and three mitochondrial genes as previously described¹⁸ using DNA extracted from a blood sample of a wild-born male *S. paradoxus* from the northern Dominican Republic (Cordillera Septentrional, Provincia de Espailat), kept at the National Zoological Park (ZOODOM) in Santo Domingo. *S. paradoxus* DNA sequences were aligned (13,885 base pairs (bp) after removal of regions of ambiguous homology) to those of taxa from all extant eutherian orders of mammals¹⁸. Figure 1 depicts the phylogenetic position of solenodons relative to other eulipotyphlan insectivores (including the results of a separate analysis to place *S. cubanus*, see below). *Solenodon* grouped with eulipotyphlan insectivores with 100% maximum-likelihood bootstrap support and bayesian posterior probability (BPP) of 1.00. Putative affinities of *Solenodon* to tenrecs^{12,15,16} or to rodents¹⁹ received no support (Supplementary Information). There was high support for *Solenodon* being the most basal eulipotyphlan (95% maximum-likelihood bootstrap support; BPP of 1.00). *Solenodon* had a more basal position than had been suggested by previous molecular or morphological reports, relative to talpids^{8,9} and/or to erinaceids^{9–11,17}.

We used well-established fossil dates²⁰ as minimum and maximum calibration points to estimate, using the method of Thorne-Kishino^{21,22}, the divergence date for *Solenodon* versus other placental mammals to be 76 million years (Myr) ago (95% credibility interval (CI) of 72–81 Myr ago) (Fig. 1 and Supplementary Information). The estimate for solenodon divergence (76 Myr ago) is comparable to or older than the estimated dates of some interordinal splits in mammals (for example, pangolins versus carnivores, or manatees versus elephants)²⁰, and considerably older than the basal divergence of most mammalian orders. The point estimate is 11 million years before the Cretaceous/Tertiary boundary at 65 Myr ago^{3,7}, with the 95% CI for solenodon divergence falling completely within the Mesozoic. The Mesozoic divergence date contrasts with previously reported support for Cenozoic divergence versus extant mainland forms for eight of nine distinct West Indian amphibian lineages, 67 of 68 reptile lineages, all 300–500 independent colonizations by birds, all 42 bat lineages, and the eight non-flying non-insectivore mammalian lineages⁴.

West Indian insectivores are therefore the only tetrapod lineage for which strong evidence supports Mesozoic divergence versus extant mainland forms, with the possible exceptions of the frog genus *Eleutherodactylus* and the Cuban xantusiid lizard *Cricosaura typica*^{4,7}. For the frog *Eleutherodactylus*, an intra-Antillean split within the genus has been previously dated to 70 ± 6.8 Myr ago⁴. For *Cricosaura typica* and related mainland lizards, we applied the Thorne-Kishino dating method^{21,22} to previously published sequences²³. While uncertain fossil constraints for xantusiids did not allow the definitive establishment of a Mesozoic origin for *Cricosaura* (95% CI of 57–101 Myr ago), the point estimate for the divergence of Cuban versus mainland xantusiids was 76 Myr ago (Supplementary Information).

Various biogeographic hypotheses have been proposed to account for the presence of solenodons only in the Antilles¹². These invoke vicariance (biogeographic separation caused by the tectonic motion of land masses or rising sea levels) or dispersal (for example, rafting across the sea on vegetation) or some combination

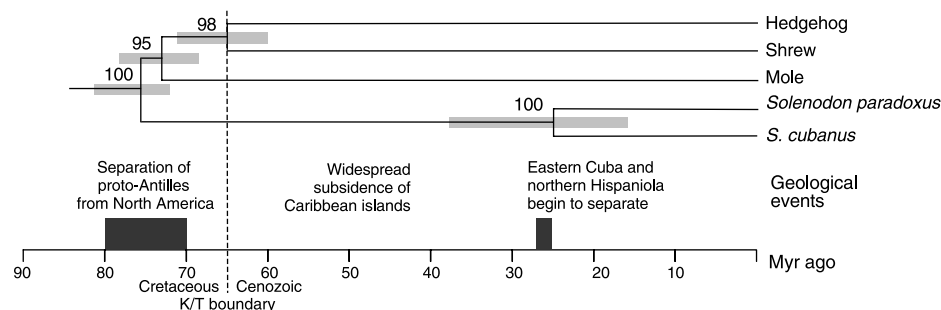


Figure 1 Phylogeny and molecular timescale for solenodons and other eulipotyphlan insectivores. The phylogeny depicted is based on analysis of a DNA sequence alignment (13.9 kb after removal of ambiguous regions) of nuclear and mitochondrial genes from *Solenodon paradoxus* and 42 other eutherian taxa²⁰, and a separate analysis of 1.6 kb of mtDNA that included *S. cubanus* (Supplementary Information). Only eulipotyphlan taxa are depicted, with nodes positioned on the basis of divergence times estimated by the Thorne–Kishino method^{21,22} using twelve fossil constraints²⁰ (Supplementary Information). Shading along each node indicates the 95% credibility interval for the

estimated divergence dates. Nodes are labelled with maximum-likelihood bootstrap support; relationships depicted were also supported by maximum-parsimony, minimum-evolution and bayesian analyses (Supplementary Information). Geological events indicated include severing of the proto-Antillean land connection to the North American mainland 70–80 Myr ago^{3,7}, the Cretaceous/Tertiary (K/T) boundary, widespread subsidence of Caribbean islands in the Tertiary^{3,4,7,26}, and the separation of eastern Cuba and northern Hispaniola, which began subsequent to 25–27 Myr ago^{3,4,7}.

of the two^{1,12,14–16}. The proto-Antillean arc, which moved north-eastward relative to the North American mainland to form part of the West Indies, was in close proximity to the mainland in the Late Cretaceous period^{3,24}. The most recent overland connection between the proto-Antilles and the North American mainland was severed 70–80 Myr ago^{3,7}. The timing of this separation is consistent with vicariance, as contemporaneous divergence dates were estimated for *Solenodon*, *Eleutherodactylus* and *Cricosaura*, even though the vicariance hypothesis must be evaluated in the context of subsequent events, notably the extraterrestrial bolide 10 km in diameter that struck the nearby Caribbean at Chicxulub 65 Myr ago^{3,4,7,25}, and the widespread subsidence of Caribbean islands in the Tertiary period^{3,4,7,26} (Fig. 1).

Dispersal during the Tertiary of a now-extinct North American mammal, such as *Apternodus* or *Centetodon*^{10,12–15}, has also been suggested to account for the presence of solenodons in the Antilles. However, a recent phylogenetic analysis has concluded that several fossil North American zalambdodonts, including *Apternodus*, were more closely related to soricids than to solenodons²⁷, which suggests that these zalambdodont taxa could not have been dispersing ancestors of solenodons. We performed a parsimony analysis of living and fossil insectivores using the same morphological data set²⁷, while employing the tree derived from our molecular analysis as a constraint. Although bootstrap values were low, solenodons were only distantly related to the soricid-fossil zalambdodont clade, or to *Centetodon* (Supplementary Information). Thus the morphological analysis does not diminish the possibility of vicariance, although it remains possible that solenodons could derive from dispersal during the Tertiary of a now-extinct North American mammal.

Individuals of *S. cubanus* are only very rarely caught, and the species has been considered extinct at various times during the past century. We used three museum samples of *S. cubanus* to examine molecular divergence in a 2.5-kilobase (kb) DNA fragment spanning portions of both mitochondrial ribosomal RNA genes. Aligning the sequence to the same mtDNA region of *S. paradoxus* and other eutherian mammals (1,624-bp alignment after removing regions of ambiguous homology), we estimated that the two *Solenodon* species diverged 25 Myr ago (95% CI of 16–38 Myr ago; Fig. 1 and Supplementary Information). For part of the Cenozoic, eastern Cuba and northern Hispaniola were attached, but after 27–25 Myr ago they began separating^{3,4,7} (Fig. 1). Our date estimate is consistent with intra-Antillean vicariance separating solenodons into two distinct island taxa. The large molecular genetic separation between the species is comparable to the divergence between

distinct mammalian families, for example, deer versus bovids (23 Myr ago)²⁸, dolphins versus whales (30 Myr ago)²⁰, or humans versus Old World monkeys (23 Myr ago)²⁸. While almost all authorities have classified Hispaniolan and Cuban solenodons as distinct species within a single genus, *Solenodon*¹, our results lend support to an alternative proposal that Cuban solenodons be classified in a distinct genus, *Atopogale*²⁹. The large genetic distance between *Solenodon* taxa may also raise the possibility that insectivore-grade early eutherian fossil taxa may be more disparate than suggested by morphological differences.

Solenodons today are threatened by deforestation, increasing human activity, the negative impacts of predation by introduced cats, dogs and mongooses, and possibly by competition from introduced rodents^{1,6}. They have survived human settlement of the islands of the West Indies, which probably eliminated two other solenodon species and eleven species of the extinct genus *Nesophontes*^{1,2}, possibly related to solenodons^{4,8–10,14}. Both solenodons are listed as endangered and declining in population by the IUCN Red List of threatened species, alongside 867 other threatened Caribbean taxa³⁰. Our results indicate that extinction of solenodons would represent the loss of an entire evolutionary lineage whose antiquity predates the extinction of the dinosaurs. They emphasize the urgent need for conservation measures on behalf of native West Indian wildlife. □

Methods

Solenodon paradoxus DNA was extracted from a fresh blood sample using a column-based kit (Qiagen), and amplified, sequenced and aligned as previously described¹⁸, resulting in a 13,885-bp data set. DNA was also successfully extracted from three museum samples of *S. cubanus* (of five individuals attempted) in a physically isolated ancient DNA laboratory, with some extractions repeated in different locations. Segments of the 12S, transfer RNA valine and 16S mitochondrial genes were amplified in overlapping fragments by nineteen pairs of primers.

Maximum-likelihood analyses were performed with PAUP* 4.0b10 and employed heuristic searches using a neighbour-joining starting tree and tree-bisection-reconnection branch swapping. Nonparametric maximum-likelihood bootstrap analysis was performed using 100 heuristic replicates with nearest-neighbour-interchange branch swapping. Settings for the GTR + Γ + I model of DNA sequence evolution were estimated initially using Modeltest and then optimized in PAUP* by additional heuristic searches. Bayesian phylogenetic analyses were performed using MrBayes v3.0b4. Maximum-parsimony and minimum-evolution analyses of *Solenodon* DNA sequences were also performed using PAUP*. To examine the relationship of *Solenodon* to fossil insectivores, a parsimony tree was generated using the morphological data set of ref. 27, employing a tree scaffold based on our molecular phylogeny.

To estimate divergence times, we used the Thorne–Kishino method^{21,22}, which permits multiple simultaneous constraints from the fossil record while allowing rates of molecular evolution to vary on different branches of a phylogenetic tree. Branch lengths were estimated with the *estbranches* program of ref. 21; divergence times were estimated using the program *divtimeSb*^{21,22}. Divergence dates were established using different DNA

sequence data sets for *S. paradoxus* versus other mammals, for *S. paradoxus* versus *S. cubanus*, and for *Cricosaura typica* versus other xantusiid lizards.

Received 26 November 2003; accepted 19 April 2004; doi:10.1038/nature02597.

- Ottenwalder, J. in *Biogeography of the West Indies: Patterns and Perspectives* (eds Woods, C. A. & Sergile, F. E.) 253–329 (CRC Press, Boca Raton, Florida, 2001).
- MacPhee, R. D. E., Fleming, C. & Lund, D. P. Last occurrence of the Antillean insectivora *Nesophontes*: new radiometric dates and their interpretation. *Am. Mus. Novit.* **3261**, 1–20 (1999).
- Iturralde-Vinent, M. A. & MacPhee, R. D. E. Paleogeography of the Caribbean region: implications for Cenozoic biogeography. *Bull. Am. Mus. Nat. Hist.* **238**, 1–95 (1999).
- Hedges, S. B. Historical biogeography of West Indian vertebrates. *Annu. Rev. Ecol. Syst.* **27**, 163–196 (1996).
- Eisenberg, J. F. & Gozalez Gotera, N. Observations on the natural history of *Solenodon cubanus*. *Acta Zool. Fenn.* **173**, 275–277 (1985).
- Ottenwalder, J. A. *The Distribution and Habitat of Solenodon in the Dominican Republic*. 1–128, MS thesis Univ. Florida, Gainesville (1985).
- Hedges, S. B., Hass, C. A. & Maxson, L. R. Caribbean biogeography: molecular evidence for dispersal in West Indian terrestrial vertebrates. *Proc. Natl Acad. Sci. USA* **89**, 1909–1913 (1992).
- McDowell, S. B. Jr The Greater Antillean insectivores. *Bull. Am. Mus. Nat. Hist.* **115**, 113–214 (1958).
- McKenna, M. C. & Bell, S. K. *Classification of Mammals above the Species Level* (Columbia Univ. Press, New York, 1997).
- Butler, P. M. in *The Phylogeny and Classification of the Tetrapods* (ed. Benton, M. J.) 117–141 (Oxford Univ. Press, Oxford, 1988).
- MacPhee, R. D. E. & Novacek, M. J. in *Mammal Phylogeny: Placentals* (eds Szalay, F. S., Novacek, M. J. & McKenna, M. C.) 13–31 (Springer, New York, 1993).
- Whidden, H. P. & Asher, R. J. in *Biogeography of the West Indies: Patterns and Perspectives* (eds Woods, C. A. & Sergile, F. E.) 237–252 (CRC Press, Boca Raton, Florida, 2001).
- Lillegraven, J. A., McKenna, M. C. & Krishtalka, L. Evolutionary relationships of middle Eocene and younger species of *Centetodon* (Mammalia, Insectivora, Geolabidae) with a description of the dentition of *Ankylodon* (Adapisoricidae). *Univ. Wyoming Publ.* **45**, 1–115 (1981).
- MacPadden, B. J. Rafting mammals or drifting islands? Biogeography of the Greater Antillean insectivores *Nesophontes* and *Solenodon*. *J. Biogeogr.* **7**, 11–22 (1980).
- Asher, R. J. A morphological basis for assessing the phylogeny of the “Tenrecoidea” (Mammalia, Lipotyphla). *Cladistics* **15**, 231–252 (1999).
- Hershkovitz, P. in *Evolution, Mammals, and Southern Continents* (eds Keast, A., Erk, F. C. & Glass, B.) 311–431 (State Univ. New York Press, Albany, 1972).
- Stanhope, M. J. *et al.* Molecular evidence for multiple origins of Insectivora and for a new order of endemic African insectivore mammals. *Proc. Natl Acad. Sci. USA* **95**, 9967–9972 (1998).
- Murphy, W. J. *et al.* Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348–2351 (2001).
- Emerson, G. L., Kilpatrick, C. W., McNiff, B. E., Ottenwalder, J. & Allard, M. W. Phylogenetic relationships of the order Insectivora based on complete 12S rRNA sequences from mitochondria. *Cladistics* **15**, 221–230 (1999).
- Springer, M. S., Murphy, W. J., Eizirik, E. & O’Brien, S. J. Placental mammal diversification and the Cretaceous–Tertiary boundary. *Proc. Natl Acad. Sci. USA* **100**, 1056–1061 (2003).
- Thorne, J. L., Kishino, H. & Painter, I. S. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**, 1647–1657 (1998).
- Kishino, H., Thorne, J. L. & Bruno, W. J. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* **18**, 352–361 (2001).
- Hedges, S. B. & Bezy, R. L. Phylogeny of xantusiid lizards: concern for data and analysis. *Mol. Phylogenet. Evol.* **2**, 76–87 (1993).
- Acton, G. D., Galbrun, B. & King, J. W. Paleolatitude of the Caribbean Plate since the Late Cretaceous. in *Proc. ODP Sci. Res.* (eds Leckie, R. M., Sigurdsson, H., Acton, G. D. & Draper, G.) **165**, 149–173, (2000).
- Robertson, D. S., McKenna, M. C., Toon, O. B., Hope, S. & Lillegraven, J. A. Survival in the first hours of the Cenozoic. *GSA Bull.* **116**, 760–768 (2004).
- Donnelly, T. W. in *Insects of Panama and Mesoamerica: Selected Studies* (eds Quintero, D. & Aiello, A.) 1–13 (Oxford Univ. Press, Oxford, 1992).
- Asher, R. J., McKenna, M. C., Emry, R. J., Tabrum, A. R. & Kron, D. G. Morphology and relationships of *Apternodus* and other extinct, zalamdodont placental mammals. *Bull. Am. Mus. Nat. Hist.* **217**, 1–117 (2002).
- Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).
- Cabrera, A. *Genera Mammalium: Insectivora, Galeopithecina* (Mus. Nacl. Cien. Nat., Madrid, 1925).
- International Union for the Conservation of Nature. *The 2003 IUCN Red List of Threatened Species* (<http://www.redlist.org>).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This paper is dedicated to the memory of the Cuban naturalist Felipe Poe (1799–1891); see Supplementary Table 1 for details of the samples he collected in the 1850s. We thank C. Bell, A. Brandt, J. Brucksch, D. Castillo, N. Crumpler, M. Malasky, J. Minchoff, H. Otero, K. Scott, J. Tabler & E. Teeling. For samples, we thank the Parque Zoológico Nacional (ZOO DOM) of the Dominican Republic; J. Chupasko at the Harvard Museum of Comparative Zoology; and P. Giere at the Museum für Naturkunde, Humboldt-Universität zu Berlin. This publication has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.L.R. (roca@ncicrf.gov), S.J.O.B. (obrien@ncicrf.gov) or W.J.M. (murphywi@ncicrf.gov). The sequences reported in this study are deposited under GenBank accession numbers AY530066–AY530088.

Convergence across biomes to a common rain-use efficiency

Travis E. Huxman^{1*}, Melinda D. Smith^{2,3*}, Philip A. Fay⁴, Alan K. Knapp⁵, M. Rebecca Shaw⁶, Michael E. Loik⁷, Stanley D. Smith⁸, David T. Tissue⁹, John C. Zak⁹, Jake F. Weltzin¹⁰, William T. Pockman¹¹, Osvaldo E. Sala¹², Brent M. Haddad⁷, John Harte¹³, George W. Koch¹⁴, Susan Schwinning¹⁵, Eric E. Small¹⁶ & David G. Williams¹⁷

¹Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA

²National Center for Ecological Analysis and Synthesis, Santa Barbara, California 93101, USA

³Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06511, USA

⁴Natural Resources Research Institute, Duluth, Minnesota 55811, USA

⁵Department of Biology, Colorado State University, Fort Collins, Colorado 80523, USA

⁶Department of Global Ecology, Carnegie Institution of Washington, Stanford, California 94305, USA

⁷Department of Environmental Studies, University of California, Santa Cruz, California 95064, USA

⁸Department of Biological Sciences, University of Nevada, Las Vegas, Nevada 89154, USA

⁹Department of Biological Sciences, Texas Tech University, Lubbock, Texas 79409, USA

¹⁰Ecology and Evolutionary Biology, University of Tennessee, Knoxville, Tennessee 37919, USA

¹¹Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131, USA

¹²Department of Ecology and IFEVA, Faculty of Agronomy, University of Buenos Aires, Buenos Aires C1417DSE, Argentina

¹³Energy and Resources Group, University of California, Berkeley, California 94720, USA

¹⁴Biological Sciences and Merriam-Powell Center for Environmental Research, Northern Arizona University, Flagstaff, Arizona 86011, USA

¹⁵Biosphere 2 Center, Columbia University, Oracle, Arizona 85623, USA

¹⁶Department of Geological Sciences, University of Colorado, Boulder, Colorado 80309, USA

¹⁷Renewable Resources and Botany, University of Wyoming, Laramie, Wyoming 82071, USA

*These authors contributed equally to this work

Water availability limits plant growth and production in almost all terrestrial ecosystems^{1–5}. However, biomes differ substantially in sensitivity of aboveground net primary production (ANPP) to between-year variation in precipitation^{6–8}. Average rain-use efficiency (RUE; ANPP/precipitation) also varies between biomes, supposedly because of differences in vegetation structure and/or biogeochemical constraints⁸. Here we show that RUE decreases across biomes as mean annual precipitation increases. However, during the driest years at each site, there is convergence to a common maximum RUE (RUE_{max}) that is typical of arid ecosystems. RUE_{max} was also identified by experimentally altering the degree of limitation by water and other resources. Thus, in years when water is most limiting, deserts, grasslands and forests all exhibit the same rate of biomass production per unit rainfall, despite differences in physiognomy and site-level RUE. Global climate models^{9,10} predict increased between-year variability in precipitation, more frequent extreme drought events, and changes in temperature. Forecasts of future ecosystem behaviour should take into account this convergent feature of terrestrial biomes.

There is a compelling need to understand how terrestrial ecosystems respond to precipitation and other external drivers to permit the forecasting of potential biosphere feedback to natural and anthropogenic changes in the climate system¹¹. This is especially

important given historical trends and future models of greenhouse gases, global temperature and precipitation regimes⁹. Water is a primary resource limiting terrestrial biological activity^{1–5}, particularly in arid and semi-arid regions¹², and its availability mediates the responsiveness of communities and ecosystems to global changes^{13,14}. Indeed, ANPP, a key ecosystem process, has been shown to increase across biomes with increasing mean annual precipitation (MAP)^{2,3,7,15}. However, variability in ANPP within ecosystems does not exhibit such a clear pattern, because variability often peaks at intermediate precipitation^{6,7}. This suggests differential sensitivities of ANPP to inter-annual variability in precipitation across biomes.

Life history and biogeochemical mechanisms can interact to influence the production response of terrestrial ecosystems to precipitation^{6–8}. The evolutionary history and ecological attributes of species present in the vegetation assemblage can influence production potential as a result of constraints on growth rate imposed by trade-offs with traits for stress tolerance¹⁶. For example, primary production in arid regions is constrained by generally lower plant densities and the relatively high frequency of slow-growing stress-tolerant species that are delayed in reaching their maximum growth rates until resources become abundant¹⁷. Production can also be constrained by an interaction between climatic and biogeochemical conditions, changing the relative importance of limiting resources (for example, water, soil nitrogen, soil phosphorus or light). In this case, for sites with high production potential in years with greater than average precipitation, soil nitrogen or other limiting resources might transiently limit biological activity¹⁸. These two mechanisms are likely to operate differentially across a water availability gradient, producing the following patterns: first, water-limited regions with low production potential should be relatively insensitive to inter-annual variation in precipitation^{6,17}; second, water-limited regions with relatively high production potential should be very sensitive to variation in water availability⁷; and last, mesic sites with high production potential should exhibit relatively low sensitivity to inter-annual variability in precipitation¹⁹.

We evaluated relationships between ANPP and precipitation (both annual values for certain years and MAP) for 14 terrestrial ecosystems in nine biomes located throughout North and South America (Supplementary Information) to quantify the sensitivity (change in ANPP divided by change in precipitation) of different ecosystems to variation in precipitation. We chose ecosystems varying by an order of magnitude in annual rainfall, spanning xeric to mesic biomes, in which the relative importance of precipitation as a limiting variable might change through time. The selected data sets were additionally limited to locations where sufficient, inter-annual records of precipitation (PTT) and ANPP could be obtained. We contrasted ANPP/precipitation relationships across and within biomes to identify potential mechanisms underlying variation in ecosystem sensitivity to precipitation, and to build on our mechanistic knowledge of precipitation effects on ecosystem processes.

When evaluated across all sites and years, ANPP increased with PTT (Fig. 1a). However, there was substantial variation in sensitivity relationships between sites. In general, the greatest slopes of ANPP and precipitation occurred at the driest sites (JRN, KNZ, RV, SEV and SGS; see Methods for site abbreviations), and the lowest (or even negative) slopes occurred at the most mesic sites (AND, BCI, HBR and HFR; Fig. 1a). To some degree, this varying sensitivity reflects differences in climatic controls on ANPP between xeric and mesic biomes. Indeed, stepwise multiple regression analysis of ANPP using annual precipitation, growing season maximum temperature ($T_{\max}$), precipitation coefficient of variance and seasonality, and ANPP in the previous year indicated that ANPP at the most productive sites (more than $800 \text{ g m}^{-2} \text{ yr}^{-1}$) was more strongly correlated with $T_{\max}$ and production in the previous year

than with annual precipitation, whereas annual precipitation remained the best correlate of ANPP at the least productive sites (less than $500 \text{ g m}^{-2} \text{ yr}^{-1}$; see Supplementary Information).

The variation in the sensitivities of ANPP to precipitation with low to high MAP across the range of biomes is consistent with the hypotheses that life history and biogeochemical mechanisms can explain how ecological systems are affected by water availability. Life history (that is, vegetation) constraints influence the impact of precipitation on biological activity in a manner that can decrease with increasing precipitation, whereas biogeochemical constraints (limitation of activity by resources other than water) can increase with increasing precipitation^{7,8}. At the sites with lowest MAP, high efficiency of water use associated with individual plant growth rate is translated to high efficiency of water use at the ecosystem level. In contrast, at sites with high MAP, selection has favoured plants with high growth rates and competitive abilities for other resources rather than high efficiency of water use. The result is less effective water use by mesic vegetation; consequently, other resources such as nitrogen and light will influence ANPP more strongly. However, both in locations with high MAP and in those with low MAP, water availability is tightly linked to biogeochemical constraints through mineralization processes and leaching²⁰. Precipitation affects both nutrient availability through its effects on microbial activity and

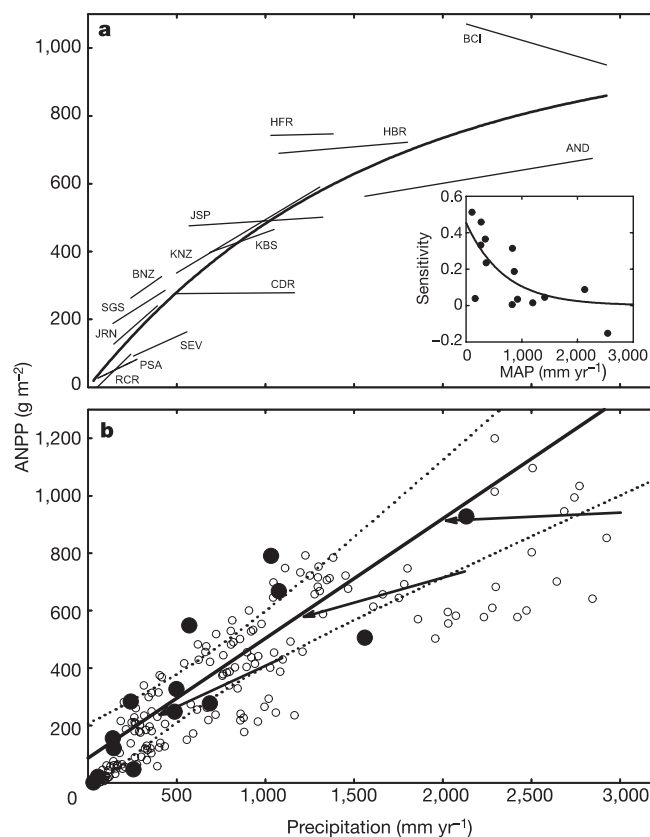


Figure 1 Between-year variation in production across a precipitation gradient and a maximum rain-use efficiency. **a**, Plot of ANPP against PPT for 14 sites (see Methods for abbreviations). Multi-year data give site-specific relationships by using linear regression (see Supplementary Information). The overall relationship (bold line) derives from data from all sites: $\text{ANPP} = 1011.7 \times (1 - \exp(-0.0006 \times \text{precipitation}))$; $r^2 = 0.77$; $P < 0.001$. The inset shows the site-level slopes (ANPP plotted against precipitation) as a function of MAP: $\text{ANPP} = 0.388 \times (1 - \exp(-0.0022 \times \text{precipitation}))$; $r^2 = 0.51$; $P < 0.001$. **b**, An overall RUE_{max} derived from the slope of the minimum precipitation and the corresponding ANPP for all sites (solid line): $\text{ANPP} = 86.1 + 0.42 \times \text{PTT}_{\min}$. Closed circles, minima; open circles, remaining data; dotted lines, 95% confidence intervals. Arrows show average slopes for sites with low, medium and high precipitation.

nutrient extraction from soils through its effects on plant growth and nutrient demand.

Variable sensitivity of ANPP to precipitation from low-production to high-production biomes reflects differences in site-level mean RUE. However, when ANPP from years with the historic minimum precipitation were combined for all sites, a positive linear relationship emerged (Fig. 1b). Thus, when water limitations on ANPP were greatest, a common RUE_{max} estimated by the slope of the historic minimum ANPP/precipitation relationship, was found for all biomes. Sites (mostly deserts) with low production potential had a mean RUE (based on all years) close to RUE_{max} whereas high-productivity sites were characterized by mean RUE that deviated significantly from RUE_{max} . Consistent with earlier analyses⁷ was our observation that intermediate sites (mostly grasslands) were more variable in yearly patterns of RUE. These sites also converged to a common RUE_{max} when water was the primary limiting resource.

Two predictions (Fig. 2a) arise from the existence of a common RUE_{max} : first, if climate change drives precipitation below a historic minimum, ANPP will be more strongly affected than predicted from site-level (mean) RUE; and second, the removal of other resource limitations so that precipitation becomes the primary limiting resource will result in an increase in site-level RUE that approaches RUE_{max} . Thus, RUE_{max} should act as a boundary by which site-level ANPP or RUE are constrained. Existing global change manipulations in a tallgrass prairie²¹ and a Mediterranean grassland¹⁴ support these two predictions (Fig. 2). First, reductions

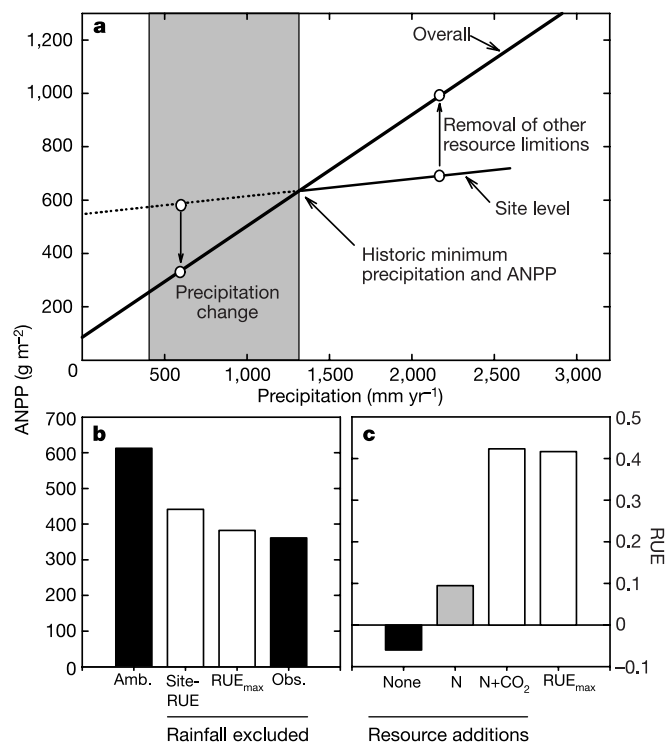


Figure 2 Hypothetical consequences of a maximum rain-use efficiency and evidence from experimental manipulations. **a**, Predictions from RUE_{max} : first, when precipitation changes below a historic minimum, RUE_{max} permits a prediction of ANPP; second, removal of resource limitations other than water will increase site RUE to RUE_{max} . **b**, Precipitation change. ANPP reduction with rain exclusion (May to October) is predicted by RUE_{max} (Konza Prairie, Kansas, USA). Filled bars, ANPP with (Obs.) and without (Amb.) rain exclusion; open bars, ANPP predicted from site-level RUE and RUE_{max} . **c**, Removal of resource limitations. The addition of resources other than water increases RUE to RUE_{max} (Jasper Ridge, California, USA). Here RUE is calculated as the change in ANPP from plots with ambient precipitation to those with water addition, for (1) no 'other' resource addition (black bar), (2) addition of soil nitrogen (N) (grey bar) and (3) addition of soil nitrogen plus CO_2 (open bar).

in ANPP with the exclusion of growing season precipitation (May to October) resulted in a substantial deviation from the predicted ANPP based on the site-level mean ANPP/precipitation relationship in tallgrass prairie. In this case, ANPP was decreased to that predicted from RUE_{max} rather than from mean RUE (Fig. 2b). Second, with the alleviation of limiting resources other than water (soil nitrogen, and soil nitrogen plus CO_2) in Mediterranean grassland, RUE increased to a value equivalent to the overall RUE_{max} (Fig. 2c). These results indicate that ANPP in terrestrial ecosystems might be fundamentally and equivalently constrained when water is most limiting, despite differences in vegetation.

The presence of a common RUE_{max} is not consistent with a simple life-history hypothesis for the control of ANPP by precipitation. Differences between species—a function of trade-offs between water-use efficiency (WUE) and growth rate—are compounded by other ecological processes at the community level; this produces divergent relationships of biomass production with water availability between the individual scale and the community scale. As a result, abiotic–biotic interactions, such as the relationship between plant-based WUE and the ecosystem transpiration/evaporation ratio, might be important in producing a common RUE_{max} rather than species traits alone. For example, at sites with low precipitation, individual plants might have higher WUE but more total precipitation might be lost to soil evaporation than to plant transpiration, decreasing system RUE below the plant-based value. In sites with higher precipitation, individual plants might have lower WUE but a greater fraction of water might move through plant transpiration, resulting in a greater balance between plant-level WUE and ecosystem RUE. The most parsimonious explanation of the divergence of local sites from the overall mean (for example, site-specific sensitivity of ANPP to precipitation) is the increasing importance of other resource limitations, and not the water-use characteristics of individual species. Because water availability has an overriding effect on all aspects of element cycling in arid lands²², nitrogen or other resources might limit production only during anomalous wet periods²³. This is not true of grasslands and forests, in which multiple factors can limit production to varying degrees²⁴.

Variation in the abundance and seasonal distribution of water availability is often used as the causal explanation for global differences in ecosystem structure and function²⁵. Here we show, through a cross-site comparison together with an examination of local processes, that the relative control of water on ANPP is a function of an overall RUE_{max} coupled with the dynamic nature of multiple limiting resources. Shifts in the timing, magnitude or variability of precipitation should have impacts on evolutionary and ecological processes that underlie this interaction between plant function, community composition and biogeochemistry^{12,26,27}. This highlights the need to understand not only the overall relationship between precipitation and biological activity but also how inter-annual variation in precipitation can affect ecosystem structure and function²⁸.

Our analysis suggests that water limitation can impose a common constraint on ANPP across diverse biomes, and that ecosystems have the same potential RUE_{max} despite differences in sensitivities of ANPP to precipitation, physiognomy, climatic history, hydrology and phylogenetic origin of representative flora. We show that differential sensitivity occurs as a result of local sites deviating from an overall RUE_{max} with the increasing influence of other resource limitations on ecological processes as precipitation increases. This suggests that biogeochemistry, rather than attributes of individual species alone, constrains community level ANPP in response to precipitation across biomes⁸. As a result, potential responses of the biosphere to changes in precipitation must be bounded by these underlying ecological constraints. Increased inter-annual variability and extreme droughts are major predictions

of global climate models^{9,10}. Thus, there might be a greater frequency of transition between ANPP limitation by water and by other limiting resources. A key result would be reductions in ANPP that were greater than expected, as well as greater variability, than that predicted by site-level models alone—even in biomes previously considered insensitive (for example, forests)—in response to future climate. □

Methods

We searched for data from a variety of sources, but included only those data sets with at least six years of concurrent measures of annual precipitation and ANPP. We assembled data from 14 sites that met these criteria, including ten US Long-Term Ecological Research (LTER) Network sites⁷, and sites in Rock Valley (RCR; desert), Nevada²⁹, Jasper Ridge Biological Preserve (JR; Mediterranean grassland), California¹⁴, Patagonia Steppe (PSA; grass/shrub steppe), Argentina³⁰, and Barro Colorado Island (BCI; tropical forest), Republic of Panama. These sites represent a broad gradient of precipitation in North and South America (105–2,542 mm MAP). The LTER sites are listed in ref. 7, and include Bonanza Creek, Alaska (BNZ), Cedar Creek, Minnesota (CDR), Harvard Forest, Massachusetts (HFR), Hubbard Brook, New Hampshire (HBF), Jornada, New Mexico (JRN), Kellogg, Michigan (KBS), Konza Prairie, Kansas (KNZ), Sevilleta, New Mexico (SEV), and Shortgrass Steppe, Colorado (SGS). We added the H.J. Andrews Experimental Forest, Oregon (AND), to this LTER data set. Data for BCI were obtained from the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC at http://www-eodis.ornl.gov/npp/npp_home.html).

Received 13 January; accepted 15 April 2004; doi:10.1038/nature02561.

- Rosenzweig, M. L. Net primary productivity of terrestrial communities: prediction from climatological data. *Am. Nat.* **102**, 67–74 (1968).
- Lieth, H. in *Primary Productivity of the Biosphere* (eds Lieth, H. & Whittaker, R. H.) 237–263 (Springer, New York, 1975).
- Webb, W. L., Lauenroth, W. K., Szarek, S. R. & Kinerson, R. S. Primary production and abiotic controls in forests, grasslands, and desert ecosystems of the United States. *Ecology* **64**, 134–151 (1986).
- Sala, O. E., Parton, W. J., Joyce, L. A. & Lauenroth, W. K. Primary production of the central grassland region of the United States. *Ecology* **69**, 40–45 (1988).
- Walter, H. Grassland, Savanne und Busch der arideren Teile Afrikas in ihrer ökologischen Bedingtheit. *Jb. Wiss. Bot.* **87**, 750–760 (1939).
- Paruelo, J. M., Lauenroth, W. K., Burke, I. C. & Sala, O. E. Grassland precipitation–use efficiency varies across a resource gradient. *Ecosystems* **2**, 64–68 (1999).
- Knapp, A. K. & Smith, M. D. Variation among biomass in temporal dynamics of aboveground primary production. *Science* **291**, 481–484 (2001).
- Veron, S. V., Paruelo, J. M., Sala, O. E. & Lauenroth, W. K. Environmental controls of primary production in agricultural systems of the Argentine Pampas. *Ecosystems* **5**, 625–635 (2002).
- IPCC. *Climate Change 2001: Synthesis Report. A Contribution of Working Groups I, II, and III to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (eds Watson, R. T. and Core Writing Team) (Cambridge Univ. Press, Cambridge, 2001).
- Easterling, D. R. *et al.* Climate extremes: Observations, modeling, and impacts. *Science* **289**, 2068–2074 (2000).
- Schimel, D. S. *et al.* Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. *Nature* **414**, 169–172 (2001).
- Noy-Meir, I. Desert ecosystems: environment and producers. *Annu. Rev. Ecol. Syst.* **4**, 51–58 (1973).
- Smith, S. D. *et al.* Elevated CO₂ increases productivity and invasive species success in an arid ecosystem. *Nature* **408**, 79–82 (2000).
- Shaw, M. R. *et al.* Grassland responses to global environmental changes suppressed by elevated CO₂. *Science* **298**, 1987–1990 (2002).
- Lauenroth, W. K. & Sala, O. E. Long-term forage production of North American shortgrass steppe. *Ecol. Appl.* **2**, 397–403 (1992).
- Grime, J. P. Evidence for existence of 3 primary strategies in plants and its relevance to ecological and evolutionary theory. *Am. Nat.* **111**, 1169–1194 (1977).
- Ehleringer, J. R. in *Terrestrial Global Productivity* (eds Roy, J., Saugier, B. & Mooney, H. A.) 345–362 (Academic, San Diego, 2001).
- Seastedt, T. R. & Knapp, A. K. Consequences of nonequilibrium resource availability across multiple time scales: the transient maxima hypothesis. *Am. Nat.* **141**, 421–433 (1993).
- Austin, A. T. & Vitousek, P. M. Nutrient dynamics on a precipitation gradient in Hawai'i. *Oecologia* **113**, 519–529 (1998).
- Chapin, F. S. III, Matson, P. A. & Mooney, H. A. *Principles of Terrestrial Ecosystem Ecology* (Springer, New York, 2002).
- Knapp, A. K. *et al.* Rainfall variability, carbon cycling, and plant species diversity in a mesic grassland. *Science* **298**, 2202–2205 (2003).
- Schlesinger, W. H. *Biogeochemistry: An Analysis of Global Change* (Academic, San Diego, 1997).
- Smith, S. D., Monson, R. K. & Anderson, J. E. *Physiological Ecology of North American Desert Plants* (Springer, New York, 1997).
- Aber, J. *et al.* Nitrogen saturation in temperate forest ecosystems. *Bioscience* **48**, 921–934 (1998).
- Ehleringer, J. R. & Mooney, H. A. in *Encyclopedia of Plant Physiology (New Series)* vol 12C (eds Lange, O. L., Nobel, P. S., Osmond, C. B. & Ziegler, H.) 205–231 (Springer, New York, 1983).
- Schlesinger, W. H. *et al.* Biological feedbacks in global desertification. *Science* **247**, 1043–1048 (1990).
- Reynolds, J. F., Virginia, R. A., Kemp, P. R., de Soya, A. G. & Tremmel, D. C. Impact of drought on desert shrubs: effects of seasonality and degree of resource island development. *Ecol. Monogr.* **69**, 69–106 (1999).
- Le Houerou, H. N., Bingham, R. L. & Skerbek, W. Relationship between the variability of primary production and the variability of annual precipitation in world arid lands. *J. Arid Environ.* **15**, 1–18 (1998).

- Turner, F. B. & Randall, D. C. Net production by shrubs and winter annuals in southern Nevada. *J. Arid Environ.* **17**, 23–36 (1989).
- Jobbagy, E. G. & Sala, O. E. Controls of grass and shrub aboveground production in the Patagonian steppe. *Ecol. Appl.* **10**, 541–549 (2000).

Supplementary Information accompanies this paper on www.nature.com/nature.

Acknowledgements We thank J. Bronstein, J. Cable, G. Davidowitz, A. Eilts, B. Enquist, D. Ignace, A. Kerkhoff, D. Potts, D. Schimel, L. Venable and M. Pavao-Zuckerman for comments on the manuscript. This work derived from an NCEAS working group, PrecipNet (principal investigator M.E.L.), supported by the National Science Foundation, the University of California, and the Santa Barbara campus. We acknowledge the support of the United States Department of Energy, the National Park Service, the National Science Foundation, and the United States Department of Agriculture.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.E.H. (huxman@email.arizona.edu) or M.D.S. (melinda.smith@yale.edu).

Harmonic-hopping in Wallacea's bats

Tigga Kingston^{1*} & Stephen J. Rossiter^{2,3*}

¹Department of Geography, Boston University, Massachusetts 02215, USA

²School of Biological Sciences, Queen Mary, University of London, London E1 4NS, UK

³School of Biological Sciences, University of Bristol, Bristol BS8 1UG, UK

* These authors contributed equally to this work

Evolutionary divergence between species is facilitated by ecological shifts, and divergence is particularly rapid when such shifts also promote assortative mating^{1–3}. Horseshoe bats are a diverse Old World family (Rhinolophidae) that have undergone a rapid radiation in the past 5 million years⁴. These insectivorous bats use a predominantly pure-tone echolocation call matched to an auditory fovea (an over-representation of the pure-tone frequency in the cochlea and inferior colliculus^{5,6}) to detect the minute changes in echo amplitude and frequency generated when an insect flutters its wings⁷. The emitted signal is the accentuated second harmonic of a series in which the fundamental and remaining harmonics are filtered out⁸. Here we show that three distinct, sympatric size morphs of the large-eared horseshoe bat (*Rhinolophus philippinensis*) echolocate at different harmonics of the same fundamental frequency. These morphs have undergone recent genetic divergence, and this process has occurred in parallel more than once⁹. We suggest that switching harmonics creates a discontinuity in the bats' perception of available prey that can initiate disruptive selection¹. Moreover, because call frequency in horseshoe bats has a dual function in resource acquisition and communication, ecological selection on frequency might lead to assortative mating and ultimately reproductive isolation and speciation, regardless of external barriers to gene flow^{1–3}.

The large-eared horseshoe bat (*Rhinolophus philippinensis*) is a rare species found from the Wallacea region of southeast Asia to northeast Australia. Observed variation in body size across its range has led to suggestions that more than one species might be present, currently recognized as subspecies¹⁰. A previous study of two size forms from Queensland, calling at 40 and 28 kHz, revealed a polyphyletic origin¹¹. We used acoustic and genetic analyses to determine the basis of phenotypic variation in this species. Bats were captured on Buton Island, southeast Sulawesi, with an additional individual from neighbouring Kabaena Island. Intensive trapping over four summers revealed low numbers of three discrete size morphs ($n = 24$) (Fig. 1).

Each morph was characterized by external features diagnostic for the species, but adults differed substantially in both forearm length (mean $\pm$ s.d. (n) for the large morph, 56.1 ± 1.5 mm (7); small morph, 47.0 ± 0.4 mm (13); intermediate (Buton island), 50.6 ± 1.4 mm (3); intermediate (Kabaena island), 48.4 mm) and body mass (large, 11.1 ± 0.5 g (7); small, 6.7 ± 0.4 g (13); Buton intermediate, 8.4 ± 0.4 g (3); Kabaena intermediate, 7.0 g). The largest morph was thus almost twice as heavy as the smallest, with a forearm difference of nearly 10 mm. No sexual dimorphism in size was observed within morphs.

Each size morph was associated with a different echolocation call frequency (large, 27.2 ± 0.2 kHz (6); small, 53.6 ± 0.6 kHz (11); Buton intermediate, 39.0 ± 0.8 kHz (3); Kabaena intermediate, 41.7 kHz). Remarkably, the emitted frequencies of the intermediate and small morphs correspond to the third and fourth harmonics of the large morph's fundamental frequency respectively, and the fundamental of the small morph matches the second harmonic of the large morph's fundamental frequency (Fig. 2). The frequency differences in these bats therefore seem to result from some form of 'harmonic-hopping'.

Switching harmonics is likely to have far-reaching consequences for the ecological interactions between morphs by creating a marked discontinuity in the bats' perception of available prey. Low frequencies have longer wavelengths, so they reflect poorly from small prey because of Rayleigh scattering^{12–14}. Consequently, the large morph is likely to encounter difficulties in detecting insects with wing lengths of less than 12.8 mm (wavelength 27.2 kHz under local conditions), and experimental evidence suggests that insects smaller than 5.0 mm might be undetectable to bats echolocating at 27 kHz^{13,14}. In contrast, prey almost half the size should still be detectable by the small morph (wavelength 6.5 mm) and possibly by the intermediate morphs (8.9 mm). However, low frequencies are less subject to environmental attenuation¹⁵, conferring greater detection ranges. We modelled prey detection distances for the three morphs from Buton for small (5 mm) and large (25 mm) prey. Backscattering reduces the target strength of small prey so severely at low frequencies that the detection distances differ little between morphs and are uniformly low (1.6–2.5 m). By comparison, atmospheric attenuation is the key factor influencing detection ranges of the large prey, with detection distances of 9.6 m for the large morph, 7.3 m for the intermediate morph and 5.5 m for the small morph. The large morph therefore samples more than five times the volume of potential 'large-prey space' than the small morph, over twice that of the intermediate, and, most importantly, more than 90 times the

volume of its own 'small prey space'. Switching harmonics could therefore create a marked shift in 'apparent resource distribution'; to the large morph, large insects appear more numerous than they are in absolute terms, and small prey (if detected) appear less abundant. Conversely, the small morph will detect almost nine times as many small insects as the large morph. This perceptual ecological discontinuity provides a novel opportunity for natural selection to initiate divergence within the population²; 54 kHz bats detect more small prey, but 27 kHz bats are at a competitive advantage in taking large insects.

Divergent ecological selection can, in turn, lead to reproductive isolation if the trait under selection also influences mate recognition and results in assortative mating^{1–3}. Selection for a clear communication channel has been implicated in acoustic divergence in cryptic bat species¹⁶, and in horseshoe bats, intraspecific communication seems to involve the temporal and phasic arrangement of the constant-frequency (CF) component of the call^{17,18}, with a repertoire of oral emissions below the fundamental¹⁹. Harmonic switching will not only shift the constant-frequency signal of each morph out of the sensitive foveal region of the other, but into areas of response minima. An acoustic 'blind spot' (auditory threshold maximum) at the fundamental frequency^{18,20} will greatly reduce the sensitivity of the small morph to the large morph's 27 kHz calls. Similarly, the large morph will be unresponsive to constant-frequency calls at 54 kHz because of a rapid decrease in sensitivity to frequencies above the auditory fovea²⁰. The large and small morphs are thus functionally deaf to each other's constant-frequency calls. Although the intermediate morph should be able to hear frequencies used by the large morph, and be heard by the small morph, in neither case is this receptivity reciprocal. Similarly, the response minima at the fundamental frequency are likely to disrupt the perception of low-frequency vocalizations among morphs. Thus, if we assume that acoustic communication is important for mate recognition, as occurs typically in bats²¹, harmonic-hopping could provide an intrinsic mechanism of instantaneous pre-mating isolation.

To test for assortative mating between morphs, we examined variation at microsatellite loci (see Supplementary Table 1) and found significant genetic differentiation between the large and small morphs ($F_{st} = 0.18$, $P < 0.001$) at a level commonly detected among island bat populations, providing strong evidence against panmixia. Despite this, the large and small morphs showed greater genetic similarity to each other than to other congeners caught at the same site (small and large morphs versus Sulawesi horseshoe bat (*R. celebensis*), $F_{st} = 0.46$, $P < 0.001$ and $F_{st} = 0.41$, $P < 0.01$, respectively; small and large morphs versus broad-eared horseshoe bat (*R. euryotis*), $F_{st} = 0.43$, $P < 0.001$ and $F_{st} = 0.40$, $P < 0.001$, respectively). The intermediate was differentiated from both the large ($F_{st} = 0.09$, $P < 0.05$) and small ($F_{st} = 0.19$, $P < 0.001$) forms, even after the Kabaena individual was excluded. Although the small samples preclude a definitive test, these data indicate that reproductive isolation between morphs has evolved recently, or might still be evolving. Assortative mating between the morphs is also reflected in differences in maximum-likelihood estimates of heterozygosity (large, 0.84; small, 0.54; intermediate, 0.74) and, in particular, in the small overlap between the range of possible values (given by the likelihood curves) of the small and large morphs (95% confidence intervals: large morph, 0.56–0.96; small, 0.25–0.77; intermediate, 0.40–0.93). Moreover, the direction of this trend suggests that the large morph is the ancestral form, because reproductive isolation by harmonic hops would create a founder effect and a consequent loss of genetic diversity in the new phonic types.

We also compared mitochondrial DNA control-region haplotypes of the three morphs from Buton with published sequences from large and intermediate morphs from Queensland, Australia¹¹. A consensus tree based on parsimony analysis revealed a reciprocally

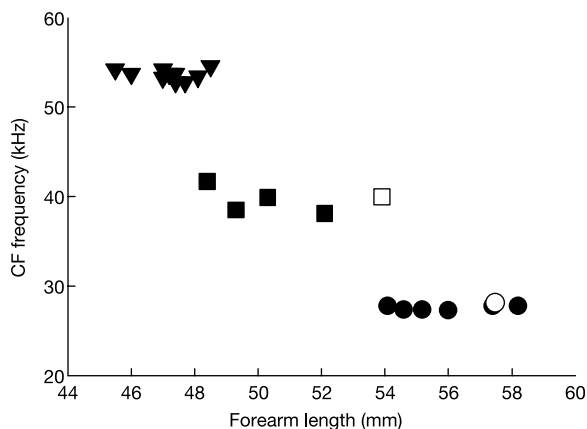


Figure 1 Relationship between constant-frequency component of calls and forearm length for three sympatric morphs of the large-eared horseshoe bat. Circle, large morph; triangle, small morph; square, intermediate morph. Filled symbols are from Sulawesi and open symbols are based on average forearm lengths³⁰ and call frequencies¹¹ from the two morphs present in Australia.

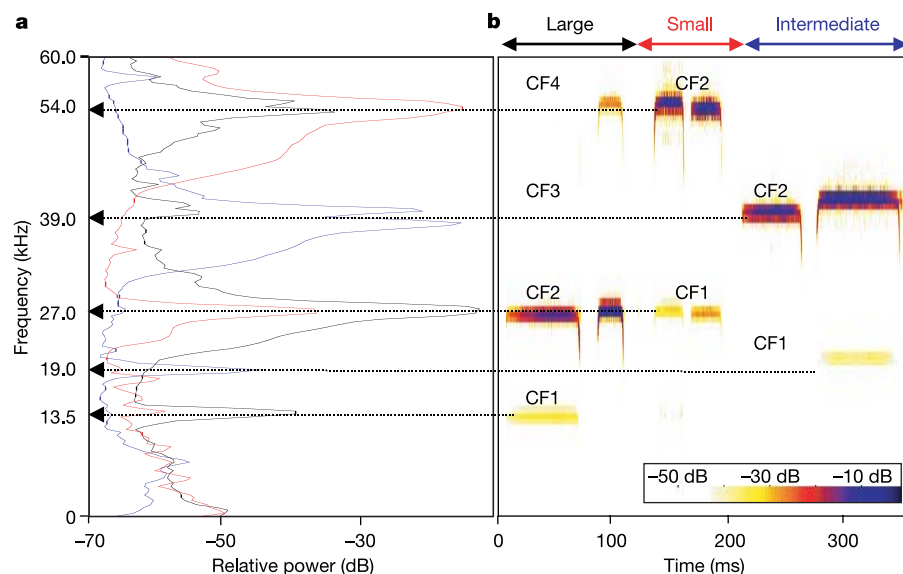


Figure 2 Echolocation calls of the three sympatric morphs of the large-eared horseshoe bat. **a**, Mean power spectra. Each includes three to eight calls per individual from 6 large morphs, 11 small morphs, and the 3 intermediate morphs from Buton. Black line, large morph; red line, small morph; blue line, Buton intermediate morph. **b**, Sonograms of two

calls from different individuals of each morph. The scale bar indicates relative power (dB). The constant-frequency harmonics of each morph are denoted by the prefix CF. CF1 is the fundamental and CF2 is the main frequency emitted. Dotted arrows indicate peak frequencies and corresponding harmonics.

monophyletic origin of both the Sulawesi and Australian morphs, with strong bootstrap support (Fig. 3). Pairwise sequence divergence (see Supplementary Table 2) was considerably lower among bats within clades (Queensland mean 0.78%, range 0.2–1.5%; Buton mean 1.06%, range 0–2.2%) than between clades (mean 7.93%, range 6.9–9.1%), indicating that distinct morphs of the large-eared horseshoe bat might have evolved more than once.

We propose that harmonic shifts provide new ecological opportunities and can promote social and thus genetic divergence regardless of the external barriers to gene flow. Remarkably, all the species of the *R. philippinensis* clade⁴, for which echolocation data are available, call at harmonics of the large morph's fundamental. *R. megaphyllus* is paraphyletic within *R. philippinensis* in Australia and echolocates at about 66 kHz¹¹ (large morph's fifth harmonic),

and *R. celebensis*, *R. virgo* and *R. borneensis* (from Sulawesi, the Philippines and Borneo, respectively) all call at about 81 kHz (sixth harmonic; ref. 22, and J. Sedlock, personal communication). We therefore suggest that harmonic shifts might have contributed to the rapid radiation of rhinolophid bats in southeast Asia, where about 30 species have originated within about 5 million years⁴. Although the origin of harmonic-hopping is unresolved, the fine tuning of the constant-frequency call to the auditory fovea necessitates a close association between the vocalization and auditory systems. We suspect that a change in cochlear dimensions initiates the switch because the functional onset and tuning of the cochlear filter are innate processes arising from postnatal maturation of the cochlea²³, which the vocalization system tracks by auditory feedback^{23,24}. Small shifts in cochlear width associated with constant-frequency shifts in horseshoe bats can occur independently of body size²². Thus, selection on cochlear width to alter the call frequency could initiate divergence and could be followed by adjustments in body size to exploit suitable prey. Such a system illustrates how, theoretically, micro-mutations could produce large-scale changes, and highlights the important function that sensory ecology is likely to have in speciation events. □

Methods

Bat capture and measurement

We undertook fieldwork on Buton Island and Kabaena Island, Sulawesi, Indonesia, from June to September over four consecutive years (2000–2003). Bats were captured in 7–14 harp traps positioned daily along forest trails within a 1-km² study plot, with the exception of one individual, which was caught with a hand net in a cave. On capture, each individual's sex, weight and forearm length were measured.

Analysis of echolocation calls

We recorded echolocation calls with a Pettersson D980 detector sampling at 400 kHz and a Pettersson D960 detector sampling at 350 kHz. Animals were held about 30 cm from the microphone for recording. Time-expanded (×10) outputs were recorded on a Sony WM-D6C Walkman cassette recorder and analysed with BatSound Pro Version 5. Power spectra were used to derive the frequency of maximum energy (kHz) and the associated harmonics for each call of each individual.

Calculation of prey size and detection distances

Mean wavelengths were calculated as the speed of sound in air divided by the call frequency, where the speed of sound is 347.65 m s⁻¹, assuming 25 °C and 80% relative humidity. Maximum target ranges were calculated as the distance at which sound losses totalled 120 dB (relative power of the emitted signal), assuming a detection threshold for *Rhinolophus* bats of 0 dB (ref. 20). The total sound loss includes absorption by the

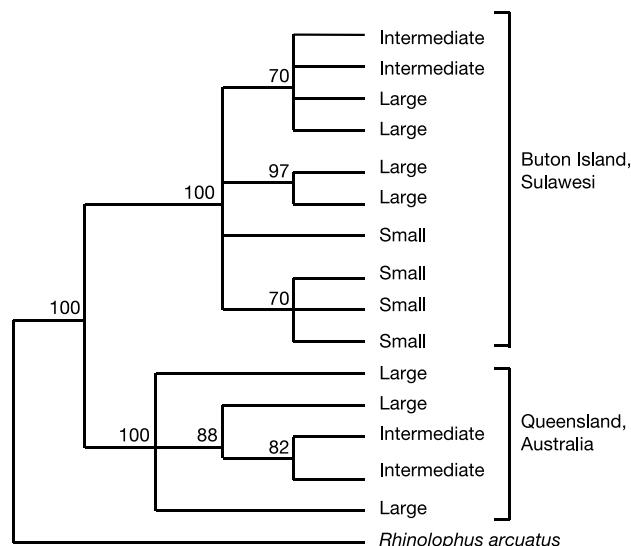


Figure 3 Consensus tree based on parsimony analysis of mtDNA haplotypes, showing phylogenetic relationships between large-eared horseshoe bat morphs sampled on Buton Island, Sulawesi, and Queensland, Australia. Bootstrap values (percentages from 1,000 replicates) are given on the nodes.

atmosphere (attenuation) and spherical spreading losses, assuming that the bat acts as a point source. The rate of atmospheric absorption is dependent on temperature, humidity and frequency and was calculated as 0.57 dB m^{-1} at 27.1 kHz, 1.09 dB m^{-1} at 39.0 kHz, 1.23 dB m^{-1} at 41.7 kHz, and 1.86 dB m^{-1} at 53.6 kHz. Spreading losses are described by the standard inverse square law for point sources. If the target is a small insect that acts as a point source it will re-radiate the incident sound back to the bat, so spreading losses are doubled¹⁵.

Total sound losses for a target at $R_2 = 2(20 \log(R_1/R_2)) + a(2(R_2 - R_1))$ (in which the two terms are spherical spreading and absorption by atmosphere, respectively) if R_1 and R_2 are two distances at a point from the original source; we follow ref. 15 in assuming that R_1 is 0.1 m in front of the bat, R_2 is the target range (distance to target) and a is absorption loss per metre. The echo strength was then adjusted for target losses, which depend on the call frequency and size of the prey. We modelled detection distance for prey with wing lengths of 25 and 5 mm with frequency-dependent target losses incorporated^{13,14}.

Tissue collection for genetic analysis

Wing membrane samples were collected from all individuals with a 3-mm biopsy punch (Stiefel Laboratories), and DNA was isolated with Qiagen DNeasy kits.

Microsatellite analysis

We genotyped *R. philippinensis* morphs (6 large, 4 intermediate, 13 small), *R. euryotis* (7 from Buton, 6 from Kabaena) and *R. celebensis* (6 from Buton) with 9–12 polymorphic microsatellite markers developed from *R. ferrumequinum* (GenBank accession numbers AF160200, AF160202, AF160207, AF160210, AF160211 (ref. 25), AJ560694, AJ560695b, AJ560698, AJ560702–AJ560704 and AJ560710 (ref. 26)). Primers were fluorescently labelled and amplified products were run on an ABI 3700 automated sequencer. Allele sizes were analysed with the software GENESCAN version 3.1 and GENOTYPER version 3.6.

Allelic diversity (number of polymorphic loci, mean number of alleles per locus) recorded for the five taxa was as follows: large morph, 8, 3.08; small morph, 8, 2.67; intermediate morph, 8, 2.25; *R. euryotis* (Buton), 9, 6.1; *R. celebensis*, 10, 4.75. We calculated Weir–Cockerham F statistics and assessed their significance with the permutation routine in the program GENETIX, which is suited to the analysis of small sample sizes. F_{is} estimates derived for each locus and for all loci together showed no significant deviation ($P > 0.05$) from Hardy–Weinberg expectations in any of the five taxa. To determine whether heterozygosity estimates differed significantly between the three *R. philippinensis* morphs, we obtained likelihood curves by following the method described in detail in ref. 27, in which the precision of the maximum-likelihood estimate of heterozygosity is reflected in the spread of the curve. To check that possible sampling of close relatives did not bias our F_{st} estimates, we examined mtDNA haplotypes and derived Queller–Goodnight pairwise coefficients of relatedness between individuals within each size class by using the program RELATEDNESS version 5.0. In only two cases of shared mtDNA haplotypes, relatedness estimates approximated to zero.

Phylogenetic inference

We amplified and sequenced 460 base pairs of the mtDNA control region with the use of the primers ThrL16272 (ref. 28) and DLH 16750 (ref. 29). Sequences were aligned with published sequences of two morphs sampled from Queensland, Australia, and an outgroup (*R. arcuatus*)¹¹. We performed phylogenetic reconstruction by maximum parsimony analysis (heuristic search) in PAUP. We derived a 50% majority-rule consensus tree and assessed the reliability of clades by bootstrapping (1,000 iterations). Pairwise divergence between all sequences was calculated (HKY85 model).

Received 3 December 2003; accepted 11 March 2004; doi:10.1038/nature02487.

- Schluter, D. Ecology and the origin of species. *Trends Ecol. Evol.* **16**, 372–380 (2001).
- Via, S. Sympatric speciation in animals: the ugly duckling grows up. *Trends Ecol. Evol.* **16**, 381–390 (2001).
- Rice, W. R. & Hostert, E. E. Perspective: Laboratory experiments on speciation: what have we learned in forty years? *Evolution* **47**, 1637–1653 (1993).
- Guillén Servent, A., Francis, C. M. & Ricklefs, R. E. in *Horseshoe Bats of the World (Chiroptera: Rhinolophidae)* (eds Corsor, G., Ujhelyi, P. & Thomas, N) xii–xxiv (Alana Books, Bishop's Castle, Shropshire, UK, 2003).
- Bruns, V. Peripheral auditory tuning for fine frequency analysis by the CF-FM bat, *Rhinolophus ferrumequinum*. I. Mechanical specializations of the cochlea. *J. Comp. Physiol.* **106**, 77–86 (1976).
- Schuller, G. & Pollak, G. D. Disproportionate frequency representation in the inferior colliculus of Doppler-compensating greater horseshoe bats, evidence for an acoustic fovea. *J. Comp. Physiol. A* **132**, 47–52 (1979).
- Schnitzler, H.-U. in *Recent Advances in the Study of Bats* (eds Fenton, M. B., Racey, P. A. & Rayner, J. M. V.) 226–243 (Cambridge Univ. Press, 1987).
- Hartley, D. J. & Suthers, R. A. The acoustics of the vocal tract in the horseshoe bat, *Rhinolophus hildebrandti*. *J. Acoust. Soc. Am.* **84**, 1201–1213 (1988).
- Schluter, D. & Nagel, L. M. Parallel speciation by natural selection. *Am. Nat.* **146**, 292–301 (1995).
- Simmons, N. B. in *Mammal Species of the World: A Taxonomic and Geographic Reference* (eds Wilson, D. E. & Reeder, D. M.) 3rd edn (Smithsonian Institution Press, Washington DC, in the press).
- Cooper, S. J. B., Reardon, T. B. & Skilins, J. Molecular systematics of Australian rhinolophid bats (Chiroptera: Rhinolophidae). *Aust. J. Zool.* **46**, 203–220 (1998).
- Pye, J. D. Is fidelity futile? The 'true' signal is illusory, especially with ultrasound. *Bioacoustics* **4**, 271–286 (1993).
- Kober, R. & Schnitzler, H.-U. Information in sonar echoes of fluttering insects available for echolocating bats. *J. Acoust. Soc. Am.* **87**, 882–896 (1990).
- Houston, R. D., Boonman, A. M. & Jones, G. in *Echolocation in Bats and Dolphins* (eds Thomas, J. A., Moss, C. F. & Vater, M.) 339–344 (Univ. of Chicago Press, 2004).

- Lawrence, B. D. & Simmons, J. A. Measurements of atmospheric attenuation at ultrasonic frequencies and the significance for echolocation by bats. *J. Acoust. Soc. Am.* **71**, 585–590 (1982).
- Jones, G. & Barlow, K. E. in *Echolocation in Bats and Dolphins* (eds Thomas, J. A., Moss, C. F. & Vater, M.) 345–349 (Univ. of Chicago Press, 2004).
- NATO Advanced Study Institute & Möhres, F. P. *Cours d'Été OTAN sur les Systèmes Sonars Animaux: Biologie et Bionique* 2 939–945 (Laboratoire de Physiologie Acoustique, Paris, 1966).
- Matsumura, S. Mother–infant communication in a horseshoe bat (*Rhinolophus ferrumequinum nippon*): vocal communication in three-week old infants. *J. Mamm.* **62**, 20–28 (1981).
- Andrews, M. M. & Andrews, P. T. Ultrasound social calls made by greater horseshoe bats (*Rhinolophus ferrumequinum*) in a nursery roost. *Acta Chiropterol.* **5**, 221–234 (2003).
- Long, G. R. & Schnitzler, H.-U. Behavioural audiograms from the bat, *Rhinolophus ferrumequinum*. *J. Comp. Physiol. A* **100**, 211–219 (1975).
- Fenton, M. B. *Communication in the Chiroptera* (Indiana Univ. Press, 1985).
- Francis, C. M. & Habersetzer, J. in *Bat Biology and Conservation* (eds Kunz, T. H. & Racey, P. A.) 169–181 (Smithsonian Institution Press, Washington DC, 1998).
- Vater, M. in *Ontogeny, Functional Ecology and Evolution of Bats* (eds Adams, R. A. & Pedersen, S. C.) 137–173 (Cambridge Univ. Press, 2000).
- Rübsamen, R. & Schäfer, M. Audio-vocal interactions during development? Vocalisation in deafened young horseshoe bats vs. audition in vocalization-impaired bats. *J. Comp. Physiol. A* **167**, 771–784 (1990).
- Rossiter, S. J., Burland, T. M., Jones, G. & Barratt, E. M. Characterization of microsatellite loci in the greater horseshoe bat *Rhinolophus ferrumequinum*. *Mol. Ecol.* **8**, 1957–1969 (1999).
- Dawson, D. A., Rossiter, S. J., Jones, G. & Faulkes, C. F. Microsatellite loci for the greater horseshoe bat, *Rhinolophus ferrumequinum* (Rhinolophidae, Chiroptera) and their cross-utility in 17 other bat species. *Mol. Ecol. Notes* **4**, 96–100 (2004).
- Nichols, R. A., Bruford, M. W. & Groombridge, J. J. Sustaining genetic variation in a small population: evidence from the Mauritius kestrel. *Mol. Ecol.* **10**, 593–602 (2001).
- Stanley, H. F. *et al.* Worldwide patterns of mitochondrial DNA differentiation in the harbour seal (*Phoca vitulina*). *Mol. Biol. Evol.* **13**, 368–382 (1996).
- Wilkinson, G. S. & Chapman, A. M. Length and sequence variation in Evening Bat D-Loop mtDNA. *Genetics* **128**, 607–617 (1991).
- Churchill, S. *Australian Bats* (New Holland, Sydney, 1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the Indonesian Institute of Science (LIPI) and the Wallacea Development Institute for granting permissions to undertake this work, and T. Coles, Boeadri and Operation Wallacea staff and volunteers for logistical support in Indonesia. We thank Rinin and Samsudin for help in the field; A. Boonman, T. Burland, D. Dawson, C. Faulkes, K. Freeman, B. Kirsten, R. Nichols, L. Pettersson and J. Storz for advice on analysis and technical support; and B. Fenton and G. Jones for helpful comments on the manuscript. This work was funded by Operation Wallacea, and microsatellite development was supported by the NERC-funded Sheffield Molecular Genetics Facility.

Authors' contributions T.K. and S.J.R. performed the acoustic and genetic analyses, respectively, and jointly undertook the fieldwork and wrote the paper. The order of authors is alphabetical.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.K. (tigga@bu.edu). Accession numbers for new and published sequences are AY568637–AY568646, and AF065069–AF065073 and AF065090 (ref. 12), respectively.

Echolocation signals reflect niche differentiation in five sympatric congeneric bat species

Björn M. Siemers & Hans-Ulrich Schnitzler

Animal Physiology, Zoological Institute, University of Tübingen, Morgenstelle 28, 72076 Tübingen, Germany

Echolocating bats can be divided into guilds according to their preferred habitat and foraging behaviour^{1–4}, which coincide with distinct adaptations in wing morphology⁵ and structure of echolocation signals⁶. Although coarse structuring of niche space between different guilds is generally accepted, it is not clear how niches differ within guilds^{7–10}, or whether there is fine-grained niche differentiation reflected in echolocation signal structure^{11,12}. Using a standardized performance test, here we show clutter-dependent differences in prey-capture success for

bats from five species of European *Myotis*. These species are morphologically similar, sympatric¹³, and all belong to the guild labelled “edge space aerial/trawling foragers”⁴. We further demonstrate a strong correlation between the prey-detection ability of the species and the respective search-call bandwidth. Our findings indicate that differences in echolocation signals contribute to within-guild niche differentiation. This is the first study relating sensory abilities of a set of potentially competing animal species to a direct measure of their respective foraging performance, suggesting an important role of sensory ecology in the structuring of animal communities.

Each of the five bat species catches flying insects or spiders on threads close to the clutter-producing foliage of vegetation^{13–18}, so they are all faced with the problem of distinguishing echoes from prey and from background or clutter^{2–4,6}. Two species also take insects from flat water surfaces by trawling, and thus have to cope with clutter echoes from drifting targets or the shore^{14,17}. Using previously published field data on foraging behaviour, diet and microhabitat use^{13–19} as well as morphometric data²⁰, we hypothesized that the five species would differ in ability to reject clutter. This, in turn, should be reflected in the minimal distance from background at which the bats could find prey by echolocation. From these data, we predicted the following order of decreasing prey-capture ability: *M. nattereri*, *M. emarginatus*, *M. mystacinus*, *M. daubentonii*, *M. dasycneme*. To test this prediction, we encouraged all species to catch prey as close as possible to a standardized clutter-producing background and measured their minimal capture distances (defined as the distance corresponding to 50% capture success).

In a flight tent, we presented mealworms to freshly captured bats at different distances to a vertical ‘clutter screen’ (polypropylene carpet with many latex-clay nubs; see Methods) that produced clutter echoes. In the dark, all bats from all of the five species located and captured silent, motionless prey close to the clutter screen. Bats also attacked rubber prey dummies lacking the odour of prey (total of 359 attacks on dummies: 39 by *M. nattereri*, 81 by *M. emarginatus*, 101 by *M. mystacinus*, 138 by *M. daubentonii*; dummies not presented to *M. dasycneme*). We conclude that visual, olfactory and passive acoustic cues were not necessary for prey detection, location and capture, all of which could be accomplished by echolocation alone. This is further corroborated by the fact that the bats of all five species consistently produced approach sequences that terminated with a ‘buzz’ (a series of very short calls with a repetition rate of about 200 Hz) typical of aerial hawking bats³. In our experiments, no bat ever captured a mealworm that was presented directly on the clutter background (Fig. 1). As in previous studies^{18,21,22}, this finding suggests that bats could not use echolocation to distinguish prey directly positioned on clutter-producing background. At distances of 25 and 50 cm almost all mealworms were captured by all bats of all five species. However, capture success differed significantly between species when prey were presented at 5 and 10 cm from clutter (Table 1) and decreased in the order of species we had predicted (Fig. 1). These differences in prey-capture ability suggest that some within-guild structuring of niche space exists.

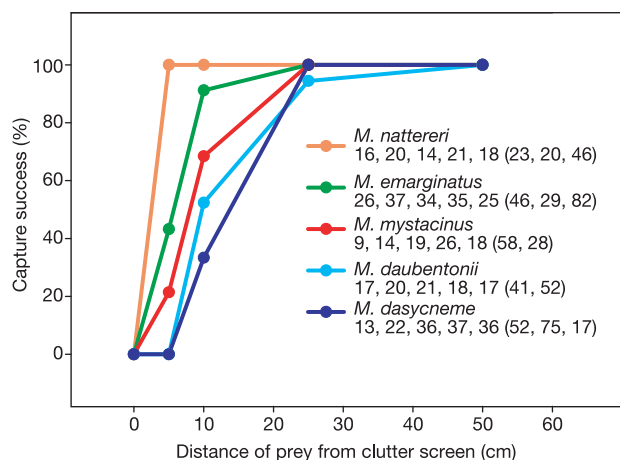


Figure 1 The capture performance of bats from five sympatric *Myotis* species searching for prey offered at different distances from a clutter screen that mimicked a vegetation edge. Species’ abilities differed at distances of 5 and 10 cm (see Table 1 for statistics). For this graph, individual performances were pooled for each species (for individual performance data see Supplementary Figure). Trials per species per distance of 0, 5, 10, 25 and 50 cm are given below the species names (parentheses enclose trials per individual). We performed a total of 569 trials.

Parameters of wing morphology, such as low wing loading or low aspect ratio, are related causally to the ability to sustain slow, manoeuvrable flight⁵, a prerequisite for consistently capturing prey close to vegetation or a clutter screen. However, although the wing loading is nearly identical in *M. emarginatus*, *M. mystacinus* and *M. daubentonii*⁵ and they indeed showed similar flight and approach behaviour in our experiments, their capture performance differed considerably. Therefore, we argue that this measured difference in capture success was caused mainly by differences in echolocation call parameters and related differences in echo-processing capabilities. To test this hypothesis we compared the measured minimal capture distances with a number of signal parameters.

The search calls used in our standardized experimental situation were steep broadband frequency-modulated (FM) signals and averaged 1.4–1.8 ms in duration (Fig. 2). The first harmonic was always most prominent. Starting, peak and terminal frequency, bandwidth and pulse interval differed between the species, whereas pulse duration showed no significant differences between the five congeners (Table 2). Capture performance was unrelated to pulse duration, peak frequency or terminal frequency (Table 2). Species with higher starting frequencies and bandwidths and shorter pulse intervals were able to capture prey closer to clutter than those with lower starting frequencies and bandwidths and higher pulse intervals (Table 2, Fig. 3). We qualitatively and quantitatively corroborated the link between bandwidth and performance by building a successful logistic model (see Supplementary Methods) with the individual performance data (see Supplementary Figure). We further examined

Table 1 Capture performance comparison for five sympatric *Myotis* species

Distance of prey from the clutter screen (cm)	Species	<i>M. emarginatus</i>	<i>M. mystacinus</i>	<i>M. daubentonii</i>	<i>M. dasycneme</i>
5	<i>M. nattereri</i>	***	***	***	***
	<i>M. emarginatus</i>		NS	**	**
	<i>M. mystacinus</i>			(*)	(*)
	<i>M. daubentonii</i>				NS
	<i>M. nattereri</i>	NS	(*)	*	***
10	<i>M. emarginatus</i>		(*)	*	***
	<i>M. mystacinus</i>			NS	(*)
	<i>M. daubentonii</i>				NS
	<i>M. nattereri</i>				

NS, not significant (probability $P > 0.05$); (*), significance lost when Bonferroni correction is applied; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

For each pair-wise comparison, all successful and unsuccessful trials of the two species were cross-tabulated and subjected to χ^2 test (degrees of freedom, 1). The P values tabulated here are Bonferroni-corrected to account for the multiple comparisons (ten comparisons per distance gives P values $\times 10$).

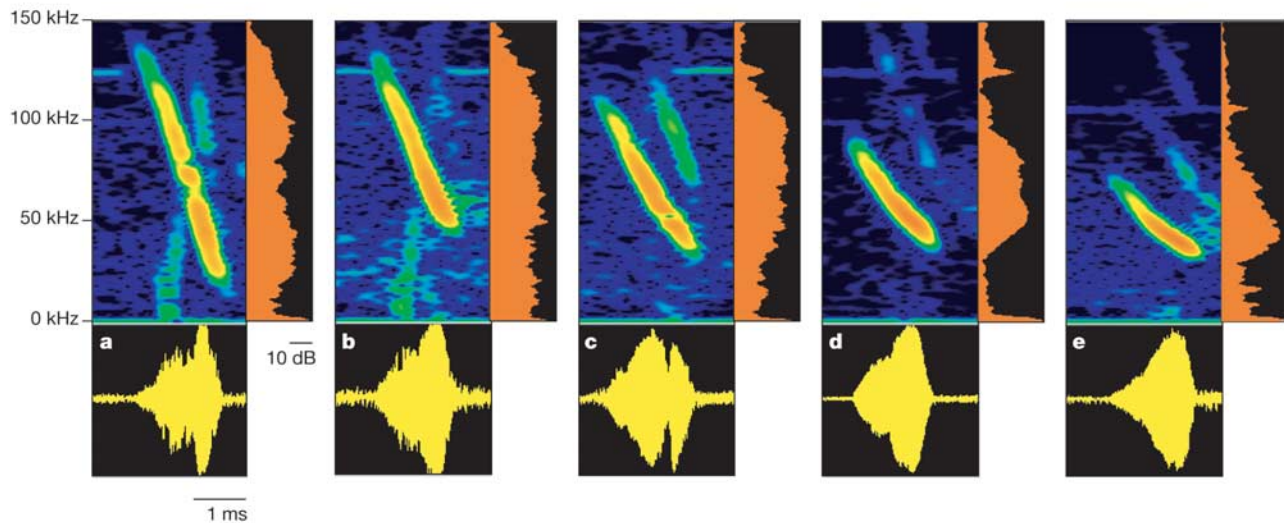


Figure 2 Representative search calls of five sympatric *Myotis* species in sonagram representation with time signal below and averaged power spectrum on the right. **a**, *Myotis nattereri*; **b**, *M. emarginatus*; **c**, *M. mystacinus*; **d**, *M. daubentonii*; and

e, *M. dasycneme*. See Methods for definition of search calls. Notches in signals (**a**, **c**) are caused by two-wavefront interference.

whether this correlation was a mere product of phylogenetic inertia by performing an independent contrast regression analysis²³ using a recent molecular phylogeny of the genus *Myotis*²⁴. Apparently, the association of broad bandwidth and good capture ability has evolved several times convergently even among the species examined, as there are two clades including a high-bandwidth species with good capture abilities and a low-bandwidth species with lower abilities (an *M. nattereri*–*M. daubentonii* clade and an *M. emarginatus*–*M. dasycneme* clade). Consequently, the correlation between bandwidth and minimal capture distance remained strong and significant for the phylogenetically corrected data set ($R^2 = 0.95$, $P = 0.0047$), with a minimal capture distance contrast slope of -0.14 with respect to contrast in bandwidth, strikingly similar to that in the non-corrected data (see Fig. 3).

When comparing the four species with low to medium capture ability, it became evident that the broadening of the bandwidth was achieved mainly by an increase in the starting frequency. The very high bandwidth of the best-performing species *M. nattereri* resulted from an additional lowering of the terminal frequency (Table 2). This may indicate that not only absolute frequency, but especially bandwidth is a crucial parameter for the improvement of prey-detection ability.

It could be argued that the bandwidth-dependent increase in range accuracy and range resolution helps bats to find prey²⁵. This explanation might be considered if the search flights were perpen-

dicular to the clutter screen, which would keep the prey echoes in front of the clutter echoes. However, the bats chose an oblique flight path and because of the width of the sonar footprint, the signals hit the screen before they hit the prey and therefore the prey echoes were buried in clutter. Hence, our experimental setting did not constitute a simple ranging task. Rather, we believe that the bats were faced with the classification problem of discriminating clutter echoes without and with prey echo. Our data show that all bats tolerated some overlap of prey echoes and background echoes.

However, until now no convincing hypothesis has been proposed to explain how bats can recognize a desired target amongst clutter, except for flower-visiting bats that use conspicuous echoes provided by the plants they pollinate²⁶. Our data may help to increase the understanding of how bats solve such complex classification problems, because they suggest that high starting frequencies and large bandwidths for good performance in such a task have adaptive value. The width of the sonar footprint on the clutter screen is reduced with increasing frequency, because of the higher directionality of the sonar beam that decreases the number of echoes from clutter at short distances. Therefore, using high-frequency signals could be a good strategy. In addition to an evolutionary increase in starting frequency, bats could achieve a behavioural reduction of the sonar footprint by keeping the search distance short²⁷. To investigate whether bats improve spatial unmasking by optimizing flight

Table 2 Search call parameters and their correlation with minimal capture distances

	Starting frequency (kHz)	Peak frequency (kHz)	Terminal frequency (kHz)	Bandwidth (kHz)	Pulse duration (ms)	Pulse interval (ms)
<i>M. nattereri</i> (3 IU, 63 calls)	135.4 ± 2.7	43.0 ± 4.1	16.2 ± 0.5	119.2 ± 2.5	1.4 ± 0.1	46.4 ± 8.9
<i>M. emarginatus</i> (3 IU, 153 calls)	133.5 ± 8.0	58.7 ± 1.6	42.2 ± 1.1	91.2 ± 8.4	1.7 ± 0.1	45.1 ± 1.8
<i>M. mystacinus</i> (2 bats, 111 calls)	107.0 ± 2.5	50.2 ± 0.5	31.6 ± 0.8	75.4 ± 3.3	1.7 ± 0.1	55.5 ± 0.9
<i>M. daubentonii</i> (2 bats, 171 calls)	90.1 ± 4.0	51.6 ± 0.5	32.9 ± 1.5	57.2 ± 2.5	1.8 ± 0.1	51.2 ± 0.2
<i>M. dasycneme</i> (3 bats, 172 calls)	73.2 ± 1.7	40.2 ± 1.1	29.4 ± 1.0	43.7 ± 2.5	1.7 ± 0.1	64.7 ± 6.6
ANOVA of call parameters (d.f. = 4)						
F-ratio	100.2	29.9	264.4	109.7	0.9	5.9
Probability, P	<0.0001	<0.0001	<0.0001	<0.0001	0.5210, NS	0.0164
Regression of minimal capture distance on call parameters (d.f. = 1)						
R ²	0.96	0.12	0.06	0.95	0.39	0.81
F-ratio	72.2	0.4	0.2	63.0	1.9	12.8
P	0.0034	0.5719, NS	0.6889, NS	0.0042	0.2602, NS	0.037

s.d., Standard deviation; d.f., degrees of freedom.

Values are shown as mean ± s.d. We used individual means to compare the call parameters between the species: 9 to 104 calls from 3 to 7 sequences for each individual (total of 670 calls). Statistical data is given under analysis of variance (ANOVA); second-order means are tabulated. To analyse the influence of the call parameters on prey-capture ability, we calculated a linear regression of minimal capture distance (species means, for data and definition see Fig. 3) on each of the call parameters (species means, see regression data).

behaviour and call directionality during both search and approach, further studies with three-dimensional analysis of flight-path and sonar-beam characteristics will be required.

We believe that an increase in bandwidth is probably advantageous because of the illumination of the sonar scene with a wider range of wavelengths: there is an 8.4-fold decrease from starting to terminal frequency in the case of the best-performing *M. nattereri*. An *M. nattereri* call that sweeps from 135 to 16 kHz contains wavelengths from 2.6 to 21.8 mm. Many arthropods (including the mealworms we used) and many reflecting background facets (leaves, grass tips and also the nubs of our clutter screen) fall within this size range. With a broadband call, many reflectors will be illuminated with wavelengths above, at and below their sizes simultaneously. Echo intensity, directionality and other reflection properties depend strongly on whether the wavelength of the incident sound is above, well below or about the same as the reflector size²⁸. Therefore, a given sonar scene may convey clear differences in the information content at frequency channels which are far apart. Bats using broadband calls presumably sample frequency-dependent differences efficiently and develop a detailed characterization of the background contours necessary for a better separation of prey from background.

Gleaning bats from the families Megadermatidae and Phyllostomidae produce broad bandwidths by using multiharmonic signals and often, but not always, rely on passive acoustic prey detection^{2,3,27}. The *Myotis* in our study as well as the palaeotropical Kerivoulinae and Murinae²⁹ (all family Vespertilionidae) achieve large call bandwidths with the first harmonic only, which may possibly prove to be a key innovation for prey detection by echolocation close to clutter.

Our data suggest that the differences in prey-detection abilities of different bats even within groups of morphologically and ecologically similar species are linked to differences in their echolocation signals and associated sensory abilities. Our data further indicate that differences in sensory abilities generate differences in prey availability for different bat species, even when foraging in exactly the same place. Therefore, differences in echolocation signals and associated sensory abilities contribute to within-guild resource partitioning. From our findings in bats, we suggest that the sensory ecologies of potentially competing species might play an important role in the structuring of animal communities. □

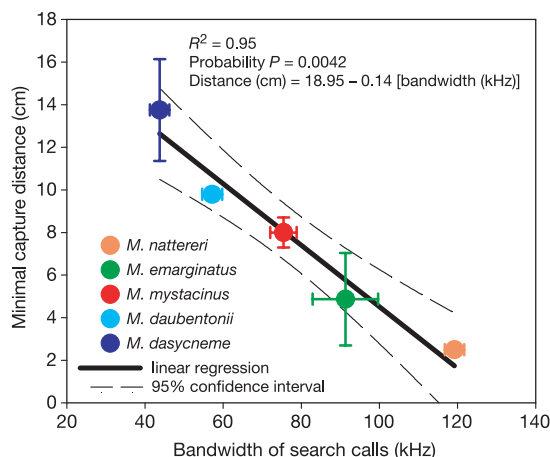


Figure 3 Correlation of 'minimal capture distance' with the bandwidth of search calls for the five sympatric *Myotis* species. We defined minimal capture distance as the distance of prey to the clutter screen that corresponded to 50% capture success. Minimal capture distance was calculated for every individual by linear interpolation from the two neighbouring distances for which performance was measured (for individual performance data see Supplementary Figure). Data are plotted as species means and error bars give the standard deviation. For number of bats see Methods.

Methods

Animals and flight tent

Bats were captured in Germany under licence from the responsible authorities (BR Hannover licence no. 503.41-22201/3; RP Karlsruhe 73c1-8852.15; RP Tübingen 73-8/8852.21; RP Stuttgart 73-8850.68-14/ Uni Tübingen; RP Freiburg Az: 73/8852.46-2) and released to the wild at the site of capture following completion of experiments. The freshly captured, experimentally naive bats were released into a mobile flight tent with a ground area of 3.5 by 7 m and about 2.5 m height, erected close to the site where the bats were captured. It had a natural light regime. We present data from three *M. dasycneme* (male), two *M. daubentonii* (male), eight *M. emarginatus* (female), two *M. mystacinus* (one male, one female) and four *M. nattereri* (one male, three female).

Behavioural experiments

Experiments were run during the natural activity period of the bats, either in complete darkness (moonless night) or with dim light (artificial or moonlight) from outside the flight tent. To assess the capture ability of the different species close to clutter-producing background, one mealworm (*Tenebrio molitor* Coleoptera) at a time was suspended on a thread (of diameter 0.1 mm) 0, 5, 10, 25 and 50 cm in front of a vertical clutter screen. A trial was scored as successful if the bat found and captured the mealworm within a 1-min time window of continuous search flight. We continuously modified the position of the mealworms in front of the clutter screen (left-right, up-down) to avoid conditioning the bats to a certain location of the screen and pseudo-randomly alternated between distances within any one session. It is important to note that we examined the bats in rather an unnatural situation, and so we did not necessarily measure the minimal detection distances bats will show in the wild, when searching for insects in front of vegetation. In natural situations, the classification problem of discriminating clutter echoes without and with prey echo will be far more complex because vegetation contours will be more irregular than our clutter screen and will deliver clutter echoes not only from the outer contour but also from the reflecting facets below³⁰.

We performed the experiments separately with only one bat flying at a time, with the exception of two *M. nattereri*, and three and four *M. emarginatus* that had to be flown together for motivational reasons. Their performances and echolocation calls were combined to form one 'individual unit' (IU) each. Thus we have presented data for 13 individuals (bats or IUs, respectively).

We used a vertical clutter screen to mimic vegetation edges in a standardized way. It consisted of a polypropylene carpet (208 cm wide, 170 cm high) with latex-clay half-spheres (nubs) of 5-mm diameter in a regular pattern with 12-mm spacing between them. When ensounded at an acute angle with a bat-like FM pulse, it returned many overlapping copies of this signal³², which is an echo complex somewhat similar to the echoes of natural vegetation³⁰.

To investigate the importance of arthropod-specific cues for prey detection, rubber dummies (electrical shrinkwrap tubing ranging from 1.6 to 12.7 mm in diameter and 1 to 47 mm in length) were offered to the bats by suspending them on nylon threads in the flight tent.

Search call recordings and sound analysis

Sequences of echolocation calls were recorded from each bat when flying in front of the clutter screen during trials in which they did not find the mealworm. We assumed that the bats were searching for (and not approaching) prey and we therefore considered these echolocation signals to be search calls. Sound recording and analysis were performed with custom-built hard- and software as described in detail elsewhere¹⁸ (256 FFT, 93.75% overlap). Sound duration and pulse interval were measured from the time signal. Starting frequency and terminal frequency were determined from a sonagram representation at 25 dB below peak frequency (that is, frequency with maximum amplitude) on the first harmonic of each signal. The -25 dB bandwidth (that is, sweep range) of each signal was computed as starting frequency minus terminal frequency.

Received 31 October 2003; accepted 5 April 2004; doi:10.1038/nature02547.

1. Aldridge, H. D. J. N. & Rautenbach, I. L. Morphology, echolocation and resource partitioning in insectivorous bats. *J. Anim. Ecol.* **56**, 763-778 (1987).
2. Neuweiler, G. Auditory adaptations for prey capture in echolocating bats. *Physiol. Rev.* **70**, 615-641 (1990).
3. Schnitzler, H.-U. & Kalko, E. K. V. Echolocation by insect-eating bats. *Bioscience* **51**, 557-569 (2001).
4. Schnitzler, H.-U., Moss, C. F. & Denzinger, A. From spatial orientation to food acquisition in echolocating bats. *Trends Ecol. Evol.* **18**, 386-394 (2003).
5. Norberg, U. M. & Rayner, J. M. V. Ecological morphology and flight in bats (Mammalia; Chiroptera): wing adaptations, flight performance, foraging strategy and echolocation. *Phil. Trans. R. Soc. Lond. B* **316**, 335-427 (1987).
6. Fenton, M. B. Foraging behaviour and ecology of animal-eating bats. *Can. J. Zool.* **68**, 411-422 (1990).
7. Findley, J. S. The structure of bat communities. *Am. Nat.* **110**, 129-139 (1976).
8. McKenzie, N. L. & Rolfe, J. K. Structure of bat guilds in the Kimberley mangroves, Australia. *J. Anim. Ecol.* **55**, 401-420 (1986).
9. Crome, F. H. J. & Richards, G. C. Bats and gaps: microchiropteran community structure in a Queensland rain forest. *Ecology* **69**, 1960-1969 (1988).
10. Sleep, D. J. H. & Brigham, R. M. An experimental test of clutter tolerance in bats. *J. Mamm.* **84**, 216-224 (2003).
11. Heller, K.-G. & von Helversen, O. Resource partitioning of sonar frequency bands in rhinolophoid bats. *Oecologia* **80**, 178-186 (1989).
12. Kingston, T., Jones, G., Zubaid, A. & Kunz, T. H. Resource partitioning in rhinolophoid bats revisited. *Oecologia* **124**, 332-342 (2000).
13. Krapp, F. in *Handbuch der Säugetiere Europas* Vol. 4/I (eds Niethammer, J. & Krapp, F.) 257-442 (Aula, Wiebelsheim, 2001).

14. Jones, G. & Rayner, J. M. V. Flight performance, foraging tactics and echolocation in free-living Daubenton's bats *Myotis daubentonii* (Chiroptera: Vespertilionidae). *J. Zool.* **215**, 113–132 (1988).
15. Krull, D., Schumm, A., Metzner, W. & Neuweiler, G. Foraging areas and foraging behavior in the notch-eared bat, *Myotis emarginatus* (Vespertilionidae). *Behav. Ecol. Sociobiol.* **28**, 247–253 (1991).
16. Schumm, A., Krull, D. & Neuweiler, G. Echolocation in the notch-eared bat, *Myotis emarginatus*. *Behav. Ecol. Sociobiol.* **28**, 255–261 (1991).
17. Britton, A. R. C., Jones, G., Rayner, J. M. V., Boonman, A. M. & Verboom, B. Flight performance, echolocation and foraging behaviour in pond bats, *Myotis dasycneme* (Chiroptera: Vespertilionidae). *J. Zool.* **241**, 503–522 (1997).
18. Siemers, B. M. & Schnitzler, H.-U. Natterer's bat (*Myotis nattereri* Kuhl, 1818) hawks for prey close to vegetation using echolocation signals of very broad bandwidth. *Behav. Ecol. Sociobiol.* **47**, 400–412 (2000).
19. Swift, S. M. & Racey, P. A. Gleaning as a foraging strategy in Natterer's bat *Myotis nattereri*. *Behav. Ecol. Sociobiol.* **52**, 408–416 (2002).
20. Fenton, M. B. & Bogdanowicz, W. Relationships between external morphology and foraging behaviour: bats in the genus *Myotis*. *Can. J. Zool.* **80**, 1004–1013 (2002).
21. Arlettaz, R., Jones, G. & Racey, P. A. Effect of acoustic clutter on prey detection by bats. *Nature* **414**, 742–745 (2001).
22. Siemers, B. M., Stitz, P. & Schnitzler, H.-U. The acoustic advantage of hunting at low heights above water: behavioural experiments on the European 'trawling' bats *Myotis capaccinii*, *M. dasycneme* and *M. daubentonii*. *J. Exp. Biol.* **204**, 3843–3854 (2001).
23. Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* 138–162 (Oxford Univ. Press, Oxford, 1991).
24. Ruedi, M. & Mayer, F. Molecular systematics of bats of the genus *Myotis* (Vespertilionidae) suggests deterministic ecomorphological convergences. *Mol. Phylog. Evol.* **21**, 436–448 (2001).
25. Simmons, J. A. & Stein, R. A. Acoustic imaging in bat sonar: Echolocation signals and the evolution of echolocation. *J. Comp. Physiol. A* **135**, 61–84 (1980).
26. von Helversen, O. & von Helversen, D. Acoustic guide in bat pollinated flower. *Nature* **398**, 759–760 (1999).
27. Schmidt, S., Hanke, S. & Pillat, J. The role of echolocation in the hunting of terrestrial prey—new evidence for an underestimated strategy in the gleaning bat, *Megaderma lyra*. *J. Comp. Physiol. A* **186**, 975–988 (2000).
28. Morse, P. M. & Ingard, K. U. *Theoretical Acoustics* (Princeton Univ. Press, Princeton, 1986).
29. Kingston, T., Jones, G., Akbar, Z. & Kunz, T. H. Echolocation signal design in Kerivoulineae and Murinae (Chiroptera: Vespertilionidae) from Malaysia. *J. Zool.* **249**, 359–374 (1999).
30. Müller, R. & Kuc, R. Foliage echoes: A probe into the ecological acoustics of bat echolocation. *J. Acoust. Soc. Am.* **108**, 836–845 (2000).

Supplementary Information accompanies this paper on www.nature.com/nature.

Acknowledgements We thank all those who assisted with fieldwork. We also thank A. Boonman, A. Denzinger, J. Ostwald, D. Menne, E. Müller, P. Pilz, M. Sánchez-Villagra and P. Stitz for discussions, H. Hartly for language assistance and B. Fenton for comments. Our research was funded by the Deutsche Forschungsgemeinschaft (DFG) and a PhD scholarship by Studienstiftung des deutschen Volkes to B.M.S.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.M.S. (bjorn.siemers@uni-tuebingen.de).

Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast

Balázs Papp^{1,2}, Csaba Pál^{1,3} & Laurence D. Hurst¹

¹Department of Biology and Biochemistry, University of Bath, BA2 7AY Bath, Somerset, UK

²Collegium Budapest (Institute for Advanced Study) Szentháromság utca 2, Budapest H-1014, Hungary

³MTA, Ecology and Theoretical Biology Research Group, Eötvös Loránd University, Pázmány Péter Sétány 1/C, Budapest H-1117, Hungary

Under laboratory conditions 80% of yeast genes seem not to be essential for viability¹. This raises the question of what the mechanistic basis for dispensability is, and whether it is the result of selection for buffering or an incidental side product. Here we analyse these issues using an *in silico* flux model^{2–5} of the yeast metabolic network. The model correctly predicts the knock-out fitness effects in 88% of the genes studied⁴ and *in vivo* fluxes.

Dispensable genes might be important, but under conditions not yet examined in the laboratory. Our model indicates that this is the dominant explanation for apparent dispensability, accounting for 37–68% of dispensable genes, whereas 15–28% of them are compensated by a duplicate, and only 4–17% are buffered by metabolic network flux reorganization. For over one-half of those not important under nutrient-rich conditions, we can predict conditions when they will be important. As expected, such condition-specific genes have a more restricted phylogenetic distribution. Gene duplicates catalysing the same reaction are not more common for indispensable reactions, suggesting that the reason for their retention is not to provide compensation. Instead their presence is better explained by selection for high enzymatic flux.

Although many single-gene deletions have negligible effects on growth rates under laboratory conditions^{1,6}, the causes and evolution of gene dispensability has remained a controversial issue^{7–9}. The capacity of organisms to compensate mutations partly stems from gene duplicates⁸, whereas alternative metabolic pathways might also have a role^{7,10–12}. The one previous systematic analysis on a eukaryotic organism¹³ used a gene's rate of evolution as a proxy for dispensability, a supposition now considered questionable¹⁴. A third possibility, and one that has received relatively little attention, is that genes only seem to be non-essential, and that they have important roles under environmental conditions yet to be replicated in the laboratory^{8,15}.

To investigate the causes of gene dispensability, the metabolic capabilities of the *Saccharomyces cerevisiae* network were calculated using flux balance analysis (FBA)¹⁶. The previously reconstructed network^{2,4} consists of 809 metabolites as nodes (including external metabolites), connected by 851 different biochemical reactions (including transport processes). The method first defines a solution space of fluxes of all metabolic reactions in the network that satisfy the governing constraints (that is, steady state of metabolites, flux capacity, direction of reactions, nutrients available in the environments; see Methods). Next, the optimal use of the metabolic network to produce major biosynthetic components for growth can be found among all possible solutions using various optimization protocols^{3,4}. The FBA and MOMA⁵ (minimization of metabolic adjustment) protocols enable us to predict the phenotypic behaviour of nutritional changes and gene deletions, along with the concomitant changes in flux distributions.

We start by asking how well the mathematical model predicts experimentally measured fluxes, and the effects of gene deletions. We then use it to address the relative importance of the suggested mechanisms for gene dispensability. Finally, we ask whether dispensability is a directly selected feature or a side consequence.

Owing to the availability of systematic knockout studies¹ and some experimentally measured fluxes under four different growth conditions¹⁷, we can directly test the predictive power of the mathematical protocol. We initiated the model to mimic the growth conditions used in these experimental studies. The model correctly predicts: (1) relative differences in flux values; (2) presence or absence of fluxes in 91–95% of the cases; (3) the fitness effects of 88% of single-gene deletions under nutrient-rich growth conditions⁴ (see Supplementary Tables S1–S3). Although the model ignores details of gene regulation, the predicted variations in the activity of metabolic pathways across environments are consistent with observations (Supplementary Tables S1 and S2; see also ref. 3). The method, although robust, is not perfect. Although the frequency of experimentally verified essential genes in the group of genes with zero predicted flux is low on rich medium, it is not zero (8.8% for genes with zero flux compared with 28.8% for the rest; $\chi^2 = 18.54$, $P < 10^{-4}$, 1 degree of freedom (d.f.)). The few essential genes in this group probably represent incomplete biochemical knowledge, missing components from the biomass equation, or pleiotropic gene functions⁴.

To investigate the possible causes of empirically observed gene dispensability, we compared the predicted and experimentally measured effects of enzyme deletions under nutrient-rich conditions (see Methods). Enzymes were classified into five mutually exclusive groupings on the basis of the presence of isoenzymes, predicted dispensability and flux distribution: (A) enzymes that are inactive under nutrient-rich conditions but active under some other environments; (B) single-copy enzymes that encode essential reactions; (C) duplicated isoenzymes that encode essential reactions; (D) single-copy enzymes that encode dispensable reactions with non-zero enzymatic flux; (E) duplicated isoenzymes that encode dispensable reactions with non-zero enzymatic flux (Fig. 1; see also Supplementary Table S4).

One possible reason that a gene might be non-essential is that its function is not required under a given circumstance. Indeed, the model predicts that a large fraction of experimentally 'verified' non-essential genes should have zero enzymatic flux under nutrient-rich conditions (68.3%). This result indicates that many enzymes make no contribution to the production of biomass components under this condition.

Can we find conditions under which the apparently non-essential genes with zero predicted flux have an important fitness contribution? As with a previous study³, we set up nine different growth conditions that might have been representative during the evolution of this species (Fig. 2), and performed enzyme-based deletion studies. Under nutrient-rich medium, the fraction of essential reactions and reactions with non-zero flux are especially low (Fig. 2). Importantly, more than half of the experimentally verified non-essential genes that are predicted to have zero flux under nutrient-rich condition appear to have non-zero flux under some other conditions (54.5%, 79 out of 145 cases). These we define as 'conditionally active' genes. This contrasts with the unconditionally active (non-zero flux under all conditions examined) genes and those for which we cannot find conditions under which they are active. These results suggest that 37–68% of the seemingly dispensable genes are environmentally specific (Supplementary Table S4).

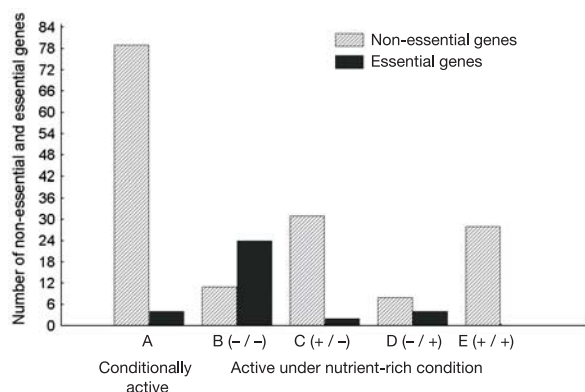


Figure 1 Number of experimentally verified essential and non-essential genes in different categories. The classes are: (A) predicted to have zero flux under nutrient-rich conditions, but non-zero flux in at least one other environment; (B) single-copy genes predicted to catalyse essential reactions; (C) duplicate genes predicted to catalyse essential reactions; (D) single-copy genes predicted to catalyse dispensable reactions; and (E) duplicate genes predicted to catalyse dispensable reactions. When comparing groups B and C, of the 68 metabolic genes that are predicted to catalyse essential reactions, 33 are known to have a duplicated isoenzyme. Only about 6% of those that have an isoenzyme are observed to be essential *in vivo*, whereas the proportion of essential genes is roughly 69% among those without an isoenzyme ($\chi^2 = 28.1$, d.f. = 1, $P < 10^{-6}$). When comparing groups B and D, of the 47 single-copy genes 35 are predicted to catalyse essential reactions whereas 12 are predicted to be dispensable. The fraction of essential genes is indeed higher in the former class (about 69% versus about 33%, $\chi^2 = 4.6$, $P < 0.05$). A plus sign indicates the presence and a minus sign the absence of isoenzymes/flux compensation.

Many of the conditionally active genes (76%) are predicted to catalyse reactions that are essential under specific conditions. It remains to be seen whether experiments will actually confirm the detailed predictions.

If the above classifications are correct, one should expect differences in the phylogenetic distribution of enzymes with unconditional and conditional activity, as the latter group can more easily be lost during evolution if the appropriate environment becomes rare. We investigated this issue using a database¹⁸ of enzymatic reactions across 133 sequenced genomes. We found that enzymes having non-zero fluxes under only a few environmental conditions tend to have a more limited phylogenetic distribution than enzymes with unconditional activity (Fig. 3).

Why is it that there are dispensable genes associated with non-zero predicted fluxes under nutrient-rich conditions? To shed light on the relative importance of the compensation mechanisms (duplication versus flux reorganization in the network), we first compared the fraction of experimentally verified essential genes between single-copy enzymes and duplicated isoenzymes. For the comparison, only genes that are predicted to encode essential reactions were considered (Fig. 1, class B and C). The low fraction of essential enzymes with isoenzymes strongly supports previous claims that dispensability partially results from redundant gene duplicates⁸. The two exceptions (failure of compensation) might be due to lack of duplicate enzyme activity in the same subcellular compartment (Supplementary Table S5). Assuming that all or none of the non-essential genes of class E are due to gene duplication rather than flux reorganization, we obtained a lower (14.6%) and upper estimate (27.8%) for the contribution of gene duplication to dispensability (Supplementary Table S4).

The ability of duplicates to buffer each other's loss may be considered a special case of a more general mode of compensation, in which the metabolic network adjusts the metabolic flux, and, in so doing, mitigates the loss of individual genes. Compensation occurs only if the original enzyme has a contribution to biomass production (non-zero flux), but the underlying reaction is dispensable for growth¹⁹ (class D and E, Fig. 1). To see the effect of flux reorganization on *in vivo* gene dispensability independent of duplicate gene copies, we compared the fraction of experimentally verified essential genes between class B and D under the assumption that essential and dispensable reactions should differ in the network's ability to compensate for their loss. Indeed, this is what we observed (Fig. 1). However, this mode of compensation can only explain 3.8–17% of gene dispensability (Fig. 1; see also Supplementary Table S4).

What factors might limit the compensatory capability of the metabolic network? Our model demonstrates that the extent of flux

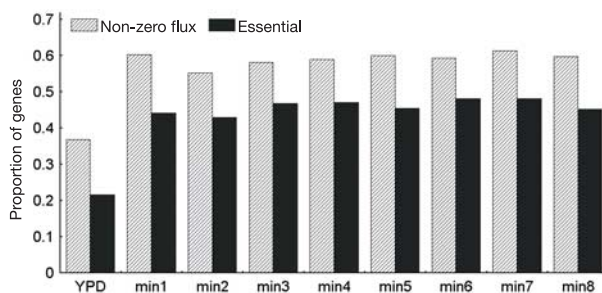


Figure 2 The proportion of genes predicted to have non-zero flux and to be essential under different growth conditions. Single-enzyme knockouts were simulated under nine different growth conditions listed below. The total number of investigated enzymes was 310 for all conditions (isoenzymes were counted only once). Environments were: YPD, rich glucose, low O_2 ; min1, minimal glucose, low O_2 ; min2, minimal glucose, anaerobic; min3, minimal ethanol, low O_2 ; min4, minimal acetate, low O_2 ; min5, minimal glucose, carbon limited; min6, minimal glucose, nitrogen limited; min7, minimal glucose, phosphate limited; min8, minimal glucose, sulphur limited.

reorganization positively correlates with the predicted fitness effect of the compensated knockout (Supplementary Fig. S1), suggesting that the yeast metabolic network has difficulties in tolerating large flux reorganization.

Although it is clear that the presence of isoenzymes has a large effect on gene dispensability, is it likely that dispensability has evolved to enable such compensation or might it instead be a side product? In the former case, one would expect gene duplicates to be preferentially maintained if they specify a crucial function to provide a shield against intracellular noise²⁰. If it were so, then one would expect the most important reactions of the network to be under the control of isoenzymes. In contrast to expectations, essential reactions are not more likely to be catalysed by isoenzymes compared to non-essential reactions (Supplementary Table S6). One possible alternative explanation for the maintenance of isoenzymes is that selection favours enhanced dosage of the same product to provide high enzymatic flux²¹. We found strong support for this theory: the average predicted flux of reactions catalysed by isoenzymes is higher under all conditions than that of reactions catalysed by single-copy enzymes (Supplementary Table S7).

Although many suggest that the high number of dispensable genes is evidence for the selection of robustness to perturbation, our results support a different conclusion. For the most part, knockout studies performed under nutrient-rich conditions provide a substantial underestimate of the number of genes that are essential under some environmental conditions (Fig. 2). Moreover, non-essential genes may make small but significant contributions to fitness even under routine growth conditions, but the effects are not large enough to be detected^{8,15}. Of those that seem to be truly dispensable (non-zero flux and viable knockout), at least in the case of gene duplicates, the dispensability is better explained as a side consequence, rather than the result of selection to favour resilience. These results, along with previous studies^{21,22}, indicate that the dosage requirements have an important influence on the evolutionary maintenance of gene duplicates in yeast.

Is it likely that environmental specificity explains much of the apparent dispensability seen in other organisms? Recent systematic deletion studies^{1,23–25} suggest that despite the apparent differences in metabolic complexity and the extent of gene duplication across free-living bacterial and eukaryotic species, the fraction of essential genes under a given laboratory condition is generally low, in the range of 7–19%. In contrast to these low figures in free-living species, the fraction of essential genes is 55–73% in the *Mycoplasma genitalium* genome²⁶. This is not simply due to a rarity of gene duplicates. We suggest that, being a parasite with strict host and tissue specificity, *M. genitalium* should have relatively few condition-specific genes. We can test this hypothesis by comparing the proportion of single-

copy genes that are non-essential in yeast and in *Mycoplasma*. In agreement with the theory, in yeast this is at least 62%, whereas this drops to 24% in *Mycoplasma*. More direct evidence comes from a data set on the growth phenotypes of mutant strains in *Escherichia coli*²⁷: most genes show severe fitness defects only under a small fraction (10%) of the 282 conditions investigated (Supplementary Fig. S2). Moreover, in agreement with the results on yeast metabolic genes, condition-specific genes of *E. coli* show limited phylogenetic distribution (Supplementary Fig. S3).

These issues are important, not least because they affect our ability to test reliably hypotheses concerning the evolution of genes and genomes. For example, the abundance of environmentally specific genes in yeast might explain why dispensability under nutrient-rich conditions only very weakly correlates with the rate of protein evolution¹⁴. □

Methods

Filtered data sets used in this study

To investigate the metabolic network we used a previously compiled list of enzymatic reactions in yeast². The metabolic reconstruction gives accurate information on the stoichiometry and direction of enzymatic reactions and on the presence of isoenzymes. Cytosolic, mitochondrial and extracellular metabolites are treated separately, and the data set also includes a list of transport reactions between compartments. Reactions catalysed by isoenzymes were considered as a single flux, eliminating duplicate reactions. For data analyses we restricted our attention to unambiguously classified enzymes; that is, those with complete EC numbers. Sequence similarity between isoenzyme pairs was computed by a pair-wise BLASTP²⁸ search (we used an *E*-value of <0.01 as a cutoff to recognize even distant duplicated isoenzymes). We checked whether the duplicated isoenzymes act in protein complexes, using both the compiled list on yeast metabolism and the MIPS CYGD²⁹ catalogue on known protein complexes, and these pairs (*N* = 21) were excluded from further analysis. Classification of the dispensability of genes on glucose-rich medium (essential versus non-essential) was as provided by the *Saccharomyces* Genome Deletion Project (http://www-sequence.stanford.edu/group/yeast_deletion_project/), which contains information on large-scale knockout studies¹. To minimize confounding factors in designation of dispensability, multienzyme polypeptides, genes participating in protein complexes (according to the MIPS CYGD catalogue of annotated complexes) and genes with overlapping reading frames were excluded from all of the analyses. The KEGG database¹⁸ was used to identify the enzymatic reactions of 133 bacterial and eukaryotic species with complete genome sequences (a filtered set of genomes consisting of only one genome per genus gives similar results).

Basic metabolic network model

Flux distribution and metabolic network capabilities were investigated by modification of a previously elaborated genome-scale metabolic flux balance model of *S. cerevisiae*^{2–4}. The model starts by specifying the mass balance constraints around intracellular metabolites. These constraints specify a series of linear equations of individual reaction fluxes that must be fulfilled to enable steady state of metabolites. Mathematically, this is represented by $Sv = 0$, where *S* is the *m* × *n* stoichiometric matrix, with *m* as the number of metabolites, and *n* as the number of reactions. An *S*_{*ij*} element of the stoichiometric matrix represents the contribution of a *j*th reaction to metabolite *i*. The vector *v* represents the individual fluxes of the network. Besides mass balance equations, reversibility/irreversibility constraints are also imposed on individual internal fluxes (*v*_{*i*} > 0 for irreversible reactions). Import flux of external metabolites was constrained to be zero when not available under the studied environment. The system also includes a biomass reaction (with rate *v*_{growth}) that represents the relative contribution of metabolites to the cellular biomass of yeast (see Supplementary equation S1). Linear programming was used to find a particular flux distribution that maximizes *v*_{growth} under the described constraints and defined nutrient uptake rates. We used this optimal flux configuration as the wild type under the given growth conditions. We have investigated nine different environments (see Fig. 2).

Calculating the fitness effect of gene knockouts

Enzyme deletions were simulated by constraining the flux of the corresponding reactions to zero and calculating the knockout flux configuration under the assumption that knockout metabolic fluxes undergo a minimal flux redistribution with respect to the flux configuration of the wild type (minimization of metabolic adjustment, MOMA protocol⁵). Using a different optimization protocol^{3,4} gives almost exactly the same results (data not shown). Thus, calculation of knockout *v*_{growth} requires quadratic programming to find a point in flux space, which is closest to wild type. The software tool Cplex 7.5 was used to solve these linear and quadratic optimization problems. We scaled fitness relative to the wild type. Essential enzymes are defined as knockout strains having a growth rate of at most one-half of the wild type. We observed a clear bimodal distribution of knockout fitnesses: genes predicted to be non-essential have minimal or no effect on growth (Supplementary Fig. S4). If the optimization problem for a given knockout was infeasible we treated the enzyme as essential. Flux and knockout phenotype predictions were not attempted for enzymes located on dead-end pathways or for enzymes with functions not represented in the biomass equation⁴ (for example, glycoprotein, haem and chitin metabolism, transfer RNA synthetases). In the case of reactions catalysed by isoenzymes, the

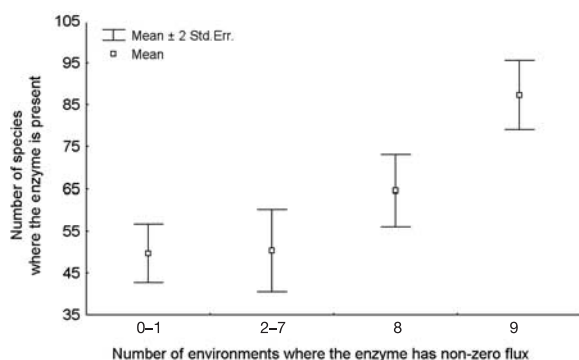


Figure 3 Relationship between phylogenetic distribution and condition specificity. Enzymes having non-zero fluxes in most of the simulated environments have wide phylogenetic distributions (analysis of variance: *F* = 17.72; d.f. = 3, 281; *P* < 10^{−9}). Data are means (square symbols) ± 2 standard errors.

duplicates were deleted to obtain predictions on the dispensability of the underlying enzymatic reaction.

Comparison of *Mycoplasma* and *Saccharomyces* genomes

We calculated the frequency of non-essential genes in the *M. genitalium* and the *S. cerevisiae* genomes (only single-copy genes were considered). Gene duplicates were identified using a BLAST protein search, with at least 25% amino acid similarity (using different thresholds do not affect our results). The list of putative essential *Mycoplasma* genes is from ref. 26. We found 1,881 out of 3,003 single-copy yeast genes that are non-essential. This figure is 83 out of 356 genes for *Mycoplasma*.

Received 14 January; accepted 30 April 2004; doi:10.1038/nature02636.

1. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
2. Forster, J., Famili, I., Fu, P., Palsson, B. O. & Nielsen, J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* **13**, 244–253 (2003).
3. Famili, I., Forster, J., Nielsen, J. & Palsson, B. O. *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl Acad. Sci. USA* **100**, 13134–13139 (2003).
4. Forster, J., Famili, I., Palsson, B. O. & Nielsen, J. Large-scale evaluation of *in silico* gene deletions in *Saccharomyces cerevisiae*. *OMICS* **7**, 193–202 (2003).
5. Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl Acad. Sci. USA* **99**, 15112–15117 (2002).
6. Steinmetz, L. M. *et al.* Systematic screen for human disease genes in yeast. *Nature Genet.* **31**, 400–404 (2002).
7. Wagner, A. Robustness against mutations in genetic networks of yeast. *Nature Genet.* **24**, 355–361 (2000).
8. Gu, Z. *et al.* Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63–66 (2003).
9. Gu, X. Evolution of duplicate genes versus genetic robustness against null mutations. *Trends Genet.* **19**, 354–356 (2003).
10. Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S. & Gilles, E. D. Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420**, 190–193 (2002).
11. Papin, J. A., Price, N. D., Wiback, S. J., Fell, D. A. & Palsson, B. O. Metabolic pathways in the post-genome era. *Trends Biochem. Sci.* **28**, 250–258 (2003).
12. Emmerling, M. *et al.* Metabolic flux responses to pyruvate kinase knockout in *Escherichia coli*. *J. Bacteriol.* **184**, 152–164 (2002).
13. Kitami, T. & Nadeau, J. H. Biochemical networking contributes more to genetic buffering in human and mouse metabolic pathways than does gene duplication. *Nature Genet.* **32**, 191–194 (2002).
14. Pál, C., Papp, B. & Hurst, L. D. Rate of evolution and gene dispensability. *Nature* **421**, 496–497 (2003).
15. Thatcher, J. W., Shaw, J. M. & Dickinson, W. J. Marginal fitness contributions of nonessential genes in yeast. *Proc. Natl Acad. Sci. USA* **95**, 253–257 (1998).
16. Edwards, J. S. & Palsson, B. O. The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl Acad. Sci. USA* **97**, 5528–5533 (2000).
17. Daran-Lapujade, P. *et al.* Role of transcriptional regulation in controlling fluxes in central carbon metabolism of *Saccharomyces cerevisiae*: a chemostat culture study. *J. Biol. Chem.* **279**, 9125–9138 (2004).
18. Nakao, M. *et al.* Genome-scale gene expression analysis and pathway reconstruction in KEGG. *Genome Inform. Ser. Workshop Genome Inform.* **10**, 94–103 (1999).
19. Schilling, C. H., Edwards, J. S., Letscher, D. & Palsson, B. O. Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol. Bioeng.* **71**, 286–306 (2000).
20. Wagner, G. P., Booth, G. & Bagheri-Chaichian, H. A population genetic theory of canalization. *Evolution* **51**, 329–347 (2000).
21. Seoighe, C. & Wolfe, K. H. Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.* **2**, 548–554 (1999).
22. Papp, B., Pál, C. & Hurst, L. D. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**, 194–197 (2003).
23. Gerdes, S. Y. *et al.* Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684 (2003).
24. Kobayashi, K. *et al.* Essential *Bacillus subtilis* genes. *Proc. Natl Acad. Sci. USA* **100**, 4678–4683 (2003).
25. Kamath, R. S. *et al.* Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237 (2003).
26. Hutchison, C. A. *et al.* Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286**, 2165–2169 (1999).
27. Glasner, J. D. *et al.* ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.* **31**, 147–151 (2003).
28. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
29. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Förster and B. Palsson for discussions and providing information on the yeast flux balance model. We also thank J. Glasner for providing growth data on the *E. coli* wild-type strain. B.P. and C.P. were supported by the Hungarian National Research Grant Foundation (OTKA).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.D.H. (l.d.hurst@bath.ac.uk).

Temporal difference models describe higher-order learning in humans

Ben Seymour¹, John P. O'Doherty¹, Peter Dayan², Martin Koltzenburg³, Anthony K. Jones⁴, Raymond J. Dolan¹, Karl J. Friston¹ & Richard S. Frackowiak^{1,5}

¹Wellcome Department of Imaging Neuroscience, 12 Queen Square, London WC1N 3BG, UK

²Gatsby Computational Neuroscience Unit, Alexandra House, 17 Queen Square, London WC1N 3AR, UK

³Institute of Child Health, University College London, 30 Guilford St, London WC1N 1EH, UK

⁴University of Manchester Rheumatic Diseases Centre, Hope Hospital, Manchester M6 8HD, UK

⁵Fondazione Santa Lucia, 00179 Rome, Italy

The ability to use environmental stimuli to predict impending harm is critical for survival. Such predictions should be available as early as they are reliable. In pavlovian conditioning, chains of successively earlier predictors are studied in terms of higher-order relationships, and have inspired computational theories such as temporal difference learning¹. However, there is at present no adequate neurobiological account of how this learning occurs. Here, in a functional magnetic resonance imaging (fMRI) study of higher-order aversive conditioning, we describe a key computational strategy that humans use to learn predictions about pain. We show that neural activity in the ventral striatum and the anterior insula displays a marked correspondence to the signals for sequential learning predicted by temporal difference models. This result reveals a flexible aversive learning process ideally suited to the changing and uncertain nature of real-world environments. Taken with existing data on reward learning², our results suggest a critical role for the ventral striatum in integrating complex appetitive and aversive predictions to coordinate behaviour.

Substantial evidence in humans and other animals has outlined a network of brain regions involved in the prediction of painful and aversive events^{3–6}. Most of this work has concentrated on its simplest realization, namely first-order pavlovian fear conditioning; however, the predictions in this paradigm are rudimentary, showing little of the complexities associated with sequences of predictors that are critical in psychological investigations of prognostication⁷. These latter studies led to a computational account called temporal difference learning^{1,8}, which has close links with methods for prediction, and optimal action selection, in engineering⁹. When applied to first-order appetitive conditioning, temporal difference learning provides a compelling account of neurophysiological data, both with respect to the phasic activity of dopamine neurons in animal studies, and with blood-oxygenation-level-dependent (BOLD) activity in human functional neuroimaging studies^{10–15}. However, beyond this simple paradigm, the utility of temporal difference models to describe learning remains largely unexplored. Here we provide a neurobiological investigation based on aversive and, importantly, sequential conditioning.

We used fMRI to investigate the pattern of brain responses in humans during a second-order pain learning task. Fourteen healthy subjects were shown two visual cues in succession, followed by a high- or low-intensity pain stimulus delivered to the left hand (Fig. 1a) (see Methods). Subjects were told that they were performing a study of reaction times and were asked to judge whether the cues appeared on the left or on the right side of a display monitor. The second cue in each sequence was fully predictive of the strength of the subsequently experienced pain; however, the first cue only allowed a probabilistic prediction. Thus, in a small percentage of

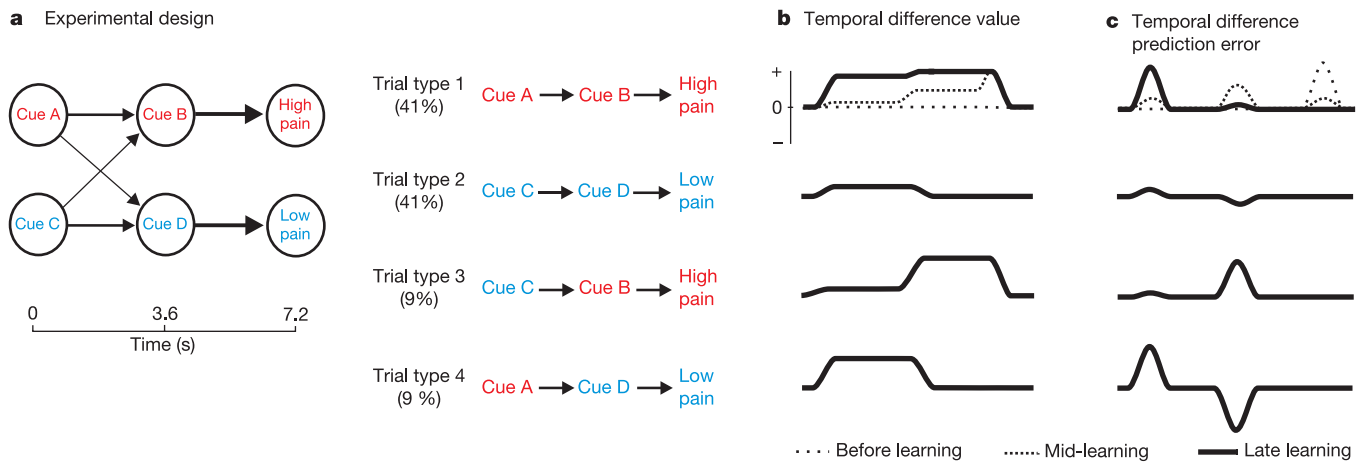


Figure 1 Experimental design and temporal difference model. **a**, The experimental design expressed as a Markov chain, giving four separate trial types. **b**, Temporal difference value. As learning proceeds, earlier cues learn to make accurate value predictions (that is, weighted averages of the final expected pain). **c**, Temporal difference prediction error;

during learning the prediction error is transferred to earlier cues as they acquire the ability to make predictions. In trial types 3 and 4, the substantial change in prediction elicits a large positive or negative prediction error. (For clarity, before and mid-learning are shown only for trial type 1.)

trials, the expectation evoked by the first cue would be reversed by the second. This allowed us to study the neural implementation of both the expectations themselves, and their reversals.

Two important aspects of most accounts of prediction learning are the predictions themselves (termed values) and the errors in those predictions⁹. Figure 1b shows the predictions associated with each of the trial types 1–4. These predictions are calculated and revised as new stimuli are presented. Figure 1c shows the associated prediction error. The nature of this signal, which treats ongoing changes in predicted values on an exact par with actual affective outcomes, has helped to explain data on dopamine cell activity. This prediction error signal drives learning by specifying how the predictions should change. In appetitive conditioning the dopamine projection to the ventral striatum is believed to be a critical substrate for this signal, although apart from theoretical speculations about opponent processing¹⁶, the equivalent for aversive conditioning is less clear. As in earlier work on appetitive conditioning, we used the temporal difference model to generate regressors based on the values and prediction errors appropriate to each individual subject¹³. Statistical parametric mapping of the regression coefficients permits identification of regions associated with, and in receipt of, information about predictions. Indeed, the temporal difference value was (negatively) correlated with the reaction times across subjects for the high-valued cues ($P < 0.001$), even when considering the second-order cue alone ($P < 0.01$). This result provides strong evidence that behavioural reinforcement occurs in a manner consistent with the temporal difference model.

The prediction error was highly correlated with activity in both the right and the left ventral putamen (Fig. 2). Correlations were also noted in the right head of the caudate, the left substantia nigra, the cerebellum (bilaterally) and the right anterior insula cortex (Fig. 2). Figure 3 shows the estimated responses in the right ventral putamen. As the most straightforward model coupling prediction error to BOLD signal would predict, positive (Fig. 3a) and negative (Fig. 3b) prediction errors at various times in the trial are clearly represented, as is the biphasic form of the prediction error (Fig. 3c).

We also investigated the representation of the temporal difference value (combining the predicted and the actual pain value, for reasons of analysis) by including the temporal difference value term in our regression model. This revealed correlated activity in the right anterior insula cortex (Fig. 4a). The estimated response is illustrated in Fig. 4c. The importance of this structure in pain learning has previously been identified⁵. In addition, we found

temporal difference value-related responses in the brainstem (Fig. 4b). Precise anatomical localization of brainstem activation is difficult with standard neuroimaging, although we note a consistency with the probable location of the dorsal raphe nucleus. We also observed temporal difference value-related responses in the anterior cingulate cortex and the right amygdala, which did not survive statistical correction for multiple comparisons.

The striking resemblance between the BOLD signal in the ventral putamen and the temporal difference prediction error (Fig. 3) offers powerful support for the temporal difference model, in particular because this is a second-order paradigm. Other dynamic models of

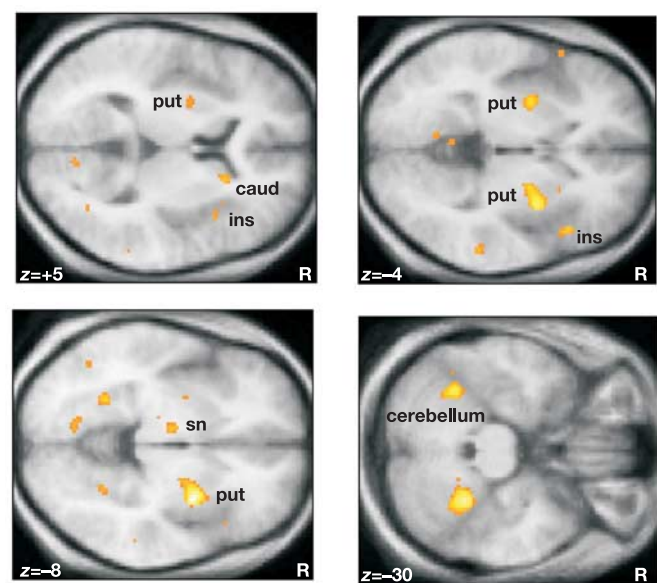


Figure 2 Temporal difference prediction error (statistical parametric maps). Areas coloured yellow/orange show significant correlation with the temporal difference prediction error. Yellow represents the greatest correlation. Peak activations (MNI coordinates and statistical z scores) are: right ventral putamen (put; (32, 0, -8), $z = 5.38$); left ventral putamen (put; (-30, -2, -4), $z = 3.93$); right head of caudate (caud; (18, 20, 6), $z = 3.75$); left substantia nigra (sn; (-10, -10, -8), $z = 3.52$); right anterior insula (ins; (46, 22, -4), $z = 3.71$); right cerebellum ((28, -46, -30), $z = 4.91$); and left cerebellum ((-34, -52, -28), $z = 4.42$). R indicates the right side.

pavlovian conditioning, such as the SOP models, do not involve this signal¹⁷. Our result adds to the growing body of neural and psychological data supporting the biological basis of temporal difference theory. In a framework called the actor–critic model for instrumental conditioning (and some variants)^{18,19}, the same prediction error signal is also used to train stimulus–response habits (called policies), ultimately leading to the choice of best possible actions²⁰. Again, this has been much more intensively studied from the perspective of appetitive than aversive conditioning. Importantly, the higher-order process demonstrated here is a crucial substrate for learning in the changing and uncertain conditions that characterize real environments, and in principle is capable of supporting complex behaviours.

Our findings add to the existing pharmacological, electrophysiological, functional imaging and clinical evidence for the involvement of the striatum in aversive processing and learning²¹. Given that the BOLD signal in the same region is correlated with the temporal difference prediction error for rewards^{13,15}, this structure may advance our understanding of precisely how aversive and appetitive information are integrated to produce motivationally appropriate behaviour in the light of (predictions of) both.

At present, the nature of the phasic aversive prediction error signal is not clear. Substantial psychological data suggest the existence of separate appetitive and aversive motivational systems that act as mutual opponents over a variety of time courses^{22,23}. Given the (not unchallenged²⁴) suggestion that dopamine neurons in the ventral tegmental area and in the substantia nigra report appetitive prediction error, it has been suggested that serotonin-producing neurons of the dorsal raphe nucleus, which send strong projections to the ventral striatum²⁵, may encode aversive prediction error¹⁶. We show prediction-related responses in an area that incorporates this nucleus. There is an active debate about the involvement of dopamine in aversive conditioning^{26,27}, and an alternative possibility is that dopamine reports both aversive and appetitive prediction errors.

Our findings have important implications for our understanding of human pain, given that substantial evidence indicates that

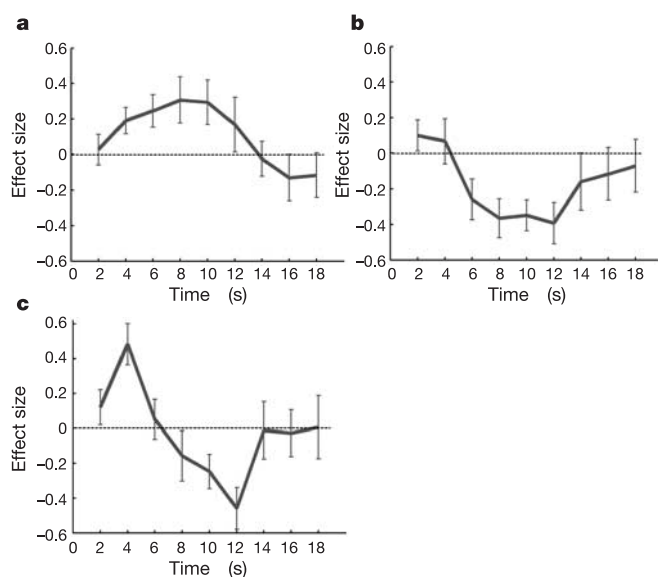


Figure 3 Temporal difference prediction error (impulse responses). Time course of the impulse response ($\pm$ s.e.m.) to higher-order prediction error in the right ventral putamen. **a**, Positive prediction error (contrast of trial types 3 and 2). **b**, Negative prediction error (contrast of trial types 4 and 1). **c**, Biphasic prediction error; positive at the first cue, becoming negative at the second (contrast of trial types 4 and 2).

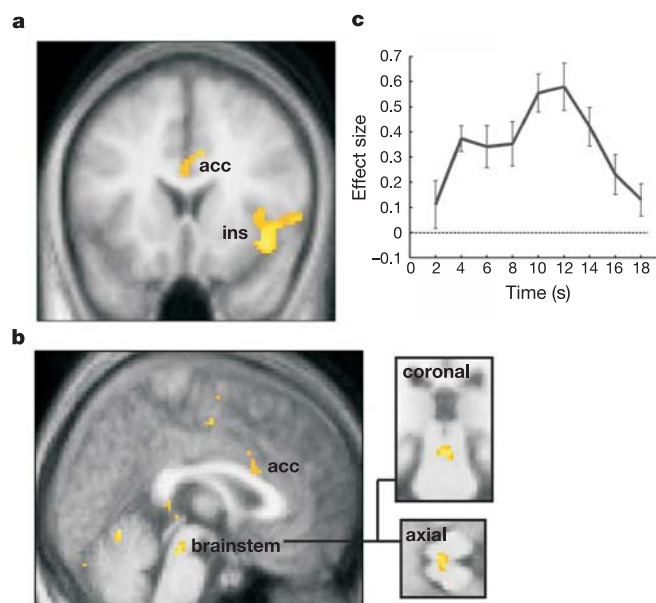


Figure 4 Temporal difference value (statistical parametric maps and impulse response in the right anterior insula). **a**, **b**, Areas showing significant correlation with the temporal difference value. Peak activations (MNI coordinates and statistical z scores) are: right anterior insula (ins; (42, 16, -14), $z = 4.16$); brainstem ((0, -28, -18), $z = 3.89$); and anterior cingulate cortex (acc; (8, 12, 32), $z = 3.82$). Coronal and axial slices of brainstem activation are shown, highlighting localization to dorsal raphe nucleus. **c**, Time course of impulse response ($\pm$ s.e.m.) in right anterior insula cortex, from contrast of trial types 1 and 2.

the experience of pain is modified by prior conditioning²⁸. Here we show regionally distinct neuronal responses, consistent with established computational processes, providing a mechanism through which the affective and motivational aspects of pain can be modulated. □

Methods

Subjects and stimuli

The study was approved by the local ethics committee. Pain was delivered to the left hand by means of two silver chloride electrodes, using a 100-Hz train of electrical pulses (4-ms square waveform, 1-s duration). A variation of current amplitude (0.5–6.0 mA) was used to deliver high- or low-intensity stimuli, set on an individual basis; mean intensity ratings were 2.9 for low intensity and 8.0 for high intensity, on a 10-point scale. Post-experiment debriefing showed no evidence of habituation or sensitization. The visual cues were abstract coloured pictures, visible on a screen by means of a head coil mirror and presented to the left or right of (and above or below) the centre. Different cues were used for the two sessions, and fully counter-balanced.

The experiment

Each subject undertook two sessions, each representing a complete learning experiment consisting of 110 trials. In each trial two cues were presented in sequence (for 3.6 s each) followed immediately by a pain stimulus. Trials were separated by a randomized variable delay of mean 5 s (range 3.5–6.5 s). The probabilistic architecture defined in Fig. 1a gives four trial types, with types 1 and 2 occurring frequently (and solely for the first ten trials). Subjects reported the position of the cue (left or right) as quickly as possible by pressing a key, and were not told that the experiment was a learning task. Post-experiment debriefing showed that no subjects were fully aware of all cue–outcome contingencies, although the influence of awareness was not specifically addressed.

Temporal difference model

The sequence of cue and pain stimuli for each subject was entered into the temporal difference learning model: TD(0)⁹ with no eligibility trace or discount factor, and on a trial basis. Each trial had three time points: first cue, second cue and pain stimulus. The six resulting states (s) were defined by the stimulus present at that time, each having a corresponding predictive value $V(s)$ (with an initial value of zero) and a return, r , representing the pain ($r = 1$ for the high intensity and $r = 0$ otherwise). At each point in time, t , the prediction error, δ , was defined as $\delta = r + V(s_t) - V(s_{t-1})$; that is, the difference between successive value predictions.

The previous state value predictions were then updated according to the algorithm $V(s) \leftarrow V(s) + \alpha\delta$, where α is the learning rate.

Data acquisition and analysis

We acquired T2*-weighted echo planar imaging (EPI) images with BOLD contrast on a 1.5 tesla Siemens Sonata magnetic resonance scanner (imaging parameters: 280 volumes; 2 mm slice thickness; 1 mm inter-slice gap; tilted plane acquisition sequence²⁹). T1-weighted structural images were co-registered with mean EPI images, and averaged across subjects to allow group level anatomical localization.

Images were analysed using the statistical parametric mapping software SPM2. Preprocessing consisted of spatial realignment, normalization to a standard EPI template, and spatial smoothing with an 8 mm (full-width at half-maximum) gaussian kernel. Images were then analysed in an event-related manner. Stimuli were encoded as δ -functions and multiplied by the temporal difference prediction error at each event provided by the computational model for each subject, and then convolved with a canonical haemodynamic response function (HRF). The parameter estimates (that is, regression coefficients) were taken to a second level random effects group analysis using a one-way analysis of variance.

Group level SPMs were initially given a threshold of $P < 0.001$, uncorrected (as shown in Figs 2 and 4). To correct for multiple comparisons, we used 8 mm radius small volume corrections (reporting family-wise error correction at $P < 0.05$) in our areas of interest, based on data from previous investigations in our laboratory³⁰ (ventral putamen, anterior insula, anterior cingulate, amygdala and cerebellum). Areas in substantia nigra (Montreal Neurological Institute (MNI) coordinates ± 10 , -14 , -10), upper brainstem (0 , -26 , -20) and dorsal striatum (± 12 , 14 , 3) were anatomically defined from our mean structural image.

We report results using a learning rate of $\alpha = 0.5$. An approximate Taylor expansion of the prediction error at $\alpha = 0.5$ and calculation of SPMs of the appropriate F -test, suggested that this was near optimal, as did results from $\alpha = 0.2$ and $\alpha = 0.8$.

Impulse responses (Figs 3 and 4b) were characterized by supplementary analysis using a flexible basis set of 2-s duration finite impulse responses on a trial basis (removing the first ten trials of early learning). Time courses shown are the averaged estimated impulse responses for each trial (with the areas of interest defined on a subject-specific basis from the original temporal difference analysis).

The sum of the temporal difference and pain values (that is, 0 or 1) were incorporated for analysis of temporal difference value (given that our design is not optimal for distinction of the two), treating prediction error and pain as effects of no interest. We applied a mask (at $P < 0.05$, uncorrected) of areas showing significant increases in activity in the cue periods from the finite impulse response analysis to ensure identification of purely predictive areas.

We normalized the reaction times (excluding those greater than 1,500 ms) using cumulative distribution functions of individually fitted gamma-distributions and regressed these values on the temporal difference value for the high pain-predicting cues. In addition to the highly significant ($P < 0.001$) regression pooled over subjects, 8 out of 14 subjects showed individually significant correlations ($P < 0.05$), but no significant fMRI differences.

Received 2 December 2003; accepted 19 April 2004; doi:10.1038/nature02581.

1. Sutton, R. S. & Barto, A. G. in *Learning and Computational Neuroscience: Foundations of Adaptive Networks* (eds Gabriel, M. & Moore, J.) 497–537 (MIT, Cambridge, Massachusetts, 1990).
2. Everitt, B. J. *et al.* Associative processes in addiction and reward. The role of amygdala–ventral striatal subsystems. *Ann. NY Acad. Sci.* **877**, 412–438 (1999).
3. LeDoux, J. Fear and the brain: where have we been, and where are we going? *Biol. Psychiatry* **44**, 1229–1238 (1998).
4. Buchel, C. & Dolan, R. J. Classical fear conditioning in functional neuroimaging. *Curr. Opin. Neurobiol.* **10**, 219–223 (2000).
5. Ploghaus, A. *et al.* Dissociating pain from its anticipation in the human brain. *Science* **284**, 1979–1981 (1999).
6. Ploghaus, A. *et al.* Learning about pain: the neural substrate of the prediction error for aversive events. *Proc. Natl Acad. Sci. USA* **97**, 9281–9286 (2000).
7. Dickinson, A. *Contemporary Animal Learning Theory* (Cambridge Univ. Press, Cambridge, UK, 1980).
8. Sutton, R. S. & Barto, A. G. Toward a modern theory of adaptive networks: expectation and prediction. *Psychol. Rev.* **88**, 135–170 (1981).
9. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT, Cambridge, Massachusetts, 1998).
10. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).
11. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
12. Suri, R. E. & Schultz, W. Temporal difference model reproduces anticipatory neural activity. *Neural Comput.* **13**, 841–862 (2001).
13. O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H. & Dolan, R. J. Temporal difference models and reward-related learning in the human brain. *Neuron* **38**, 329–337 (2003).
14. Friston, K. J., Tononi, G., Reeke, G. N. Jr, Sporns, O. & Edelman, G. M. Value-dependent selection in the brain: simulation in a synthetic neural model. *Neuroscience* **59**, 229–243 (1994).
15. McClure, S. M., Berns, G. S. & Montague, P. R. Temporal prediction errors in a passive learning task activate human striatum. *Neuron* **38**, 339–346 (2003).
16. Daw, N. D., Kakade, S. & Dayan, P. Opponent interactions between serotonin and dopamine. *Neural Netw.* **15**, 603–616 (2002).
17. Brandon, S. E., Vogel, E. H. & Wagner, A. R. Stimulus representation in SOP: I. Theoretical rationalization and some implications. *Behav. Processes* **62**, 5–25 (2003).
18. Barto, A. G., Sutton, R. S. & Anderson, C. W. Neuronlike elements that can solve difficult learning problems. *IEEE Trans. Syst. Man Cybern.* **13**, 834–846 (1983).
19. Barto, A. G., Sutton, R. S. & Watkins, C. J. C. H. in *Learning and Computational Neuroscience: Foundations of Adaptive Networks* (eds Gabriel, M. & Moore, J.) 539–602 (MIT, Cambridge, Massachusetts, 1990).

20. Barto, A. G. in *Models of Information Processing in the Basal Ganglia* (eds Houk, J. C., Davis, J. L. & Beiser, D. G.) 215–232 (MIT, Cambridge, Massachusetts, 1995).
21. Chudler, E. H. & Dong, W. K. The role of the basal ganglia in nociception and pain. *Pain* **60**, 3–38 (1995).
22. Solomon, R. L. & Corbit, J. D. An opponent-process theory of motivation. I. Temporal dynamics of affect. *Psychol. Rev.* **81**, 119–145 (1974).
23. Dickinson, A. & Dearing, M. F. in *Mechanisms of Learning and Motivation* (eds Dickinson, A. & Boakes, R. A.) 203–231 (Erlbaum, Hillsdale, New Jersey, 1979).
24. Horvitz, J. C. Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience* **96**, 651–656 (2000).
25. Azmitia, E. C. & Segal, M. An autoradiographic analysis of the differential ascending projections of the dorsal and median raphe nuclei in the rat. *J. Comp. Neurol.* **179**, 641–667 (1978).
26. Mirenowicz, J. & Schultz, W. Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature* **379**, 449–451 (1996).
27. Horvitz, J. C. Dopamine gating of glutamatergic sensorimotor and incentive motivational input signals to the striatum. *Behav. Brain Res.* **137**, 65–74 (2002).
28. Ploghaus, A., Becerra, L., Borras, C. & Borsook, D. Neural circuitry underlying pain modulation: expectation, hypnosis, placebo. *Trends Cogn. Sci.* **7**, 197–200 (2003).
29. Deichmann, R., Gottfried, J. A., Hutton, C. & Turner, R. Optimized EPI for fMRI studies of the orbitofrontal cortex. *Neuroimage* **19**, 430–441 (2003).
30. Buchel, C., Dolan, R. J., Armony, J. L. & Friston, K. J. Amygdala–hippocampal involvement in human aversive trace conditioning revealed through event-related functional magnetic resonance imaging. *J. Neurosci.* **19**, 10869–10876 (1999).

Acknowledgements We thank P. Allen and E. Featherstone for technical help. This work was funded by Wellcome Trust program grants to R.S.F., K.J.F., M.K. and R.J.D. P.D. was funded by the Gatsby Charitable foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.S. (bseymour@fil.ion.ucl.ac.uk).

Myosin-dependent junction remodelling controls planar cell intercalation and axis elongation

Claire Bertet, Lawrence Sulak & Thomas Lecuit

Laboratoire de Génétique et de Physiologie du Développement (LGPD), Institut de Biologie du Développement de Marseille (IBDM), CNRS-INSERM-Université de la Méditerranée, Campus de Luminy, case 907, 13288 Marseille cedex 9, France

Shaping a developing organ or embryo relies on the spatial regulation of cell division and shape. However, morphogenesis also occurs through changes in cell-neighbourhood relationships produced by intercalation^{1,2}. Intercalation poses a special problem in epithelia because of the adherens junctions, which maintain the integrity of the tissue. Here we address the mechanism by which an ordered process of cell intercalation directs polarized epithelial morphogenesis during germ-band elongation, the developmental elongation of the *Drosophila* embryo. Intercalation progresses because junctions are spatially reorganized in the plane of the epithelium following an ordered pattern of disassembly and reassembly. The planar remodelling of junctions is not driven by external forces at the tissue boundaries but depends on local forces at cell boundaries. Myosin II is specifically enriched in disassembling junctions, and its planar polarized localization and activity are required for planar junction remodelling and cell intercalation. This simple cellular mechanism provides a general model for polarized morphogenesis in epithelial organs.

Drosophila germ-band elongation (GBE)³ leads to an almost doubling in length of the epithelial layer that forms the thorax and abdomen of the embryo (Fig. 1a, b). This extension is associated with a mediolateral convergence of cells along the dorsoventral (D/V) axis. Neither cell divisions nor changes in cell shape contribute to anteroposterior (A/P) axis elongation³; thus, GBE is solely

associated with epithelial cell intercalation, which requires the modification of adhesion between neighbouring cells. Cell adhesion at epithelial junctions depends on homotypic binding of E-cadherin between adjacent epithelial cells. Epithelial junctions in *Drosophila* and other organisms consist of a core E-cadherin/ β -catenin/ α -catenin complex (the adherens junctions)^{4–6}, whose recruitment at the membrane is stabilized by a crosslinking network of actin filaments⁷.

To understand the cellular mechanisms of intercalation during GBE, we focused on adherens junctions containing a fully functional fusion protein between E-cadherin and green fluorescent protein (GFP) expressed at endogenous levels in early embryos⁸. The living embryos were observed by time-lapse microscopy. After 30 min, cells intercalate along the A/P axis in the ventrolateral region of the ectoderm (Fig. 1c). Cells do not delaminate, but instead remain integrated in the epithelium. Moreover, counter to a naive expectation, individual cells do not move relative to a population of more static neighbouring cells. Instead, cell intercalation seems to stem from a global rearrangement of cell junctions in a highly organized spatio-temporal pattern.

At the onset of GBE, epithelial cells form a packed hexagonal array with their boundaries either parallel to the D/V axis (referred as type 1), or at $\pm 60^\circ$ to it. When GBE begins, groups of four cells form characteristic tetrads around type 1 junctions (Fig. 1d, g, red). In this type 1 configuration, adjacent cells along the A/P axis are in contact but immediately dorsal and ventral cells are not in contact

with each other. As GBE proceeds, the $\pm 60^\circ$ junctions hardly change their orientation or average length (4 μ m). In contrast, type 1 junctions specifically shrink (average time 10 min), leading to a configuration in which the four cells of the tetrad now share equal contacts (type 2; Fig. 1e, g). Subsequently, new junctions of type 3 are built perpendicular to the old type 1 junction, resulting in effective intercalation of the cells that were dorsal and ventral (Fig. 1f, g; see also movie in Supplementary Information). This polarized pattern of junction remodelling shows that cells can distinguish between its different cell boundaries and control a specific behaviour in each of them (shrinkage or extension). These observations therefore indicate that intercalating cells are inherently polarized within the plane of the epithelium.

Observation of 314 junctions showed that 100% of them irreversibly underwent the transition type 1 to type 2 to type 3, and never the reverse transition (Fig. 1g). This leads to a relative decrease in the number of type 1 junctions and a corresponding increase in the number of type 3 junctions as intercalation proceeds (Fig. 1h). Junction remodelling occurs only in the intercalating region; that is, in the ventral and lateral ectoderm (Fig. 1a, b, grey). A group of cells that complete the transition from type 1 to type 3 elongates about 50% along the A/P axis, and correspondingly shortens along the D/V axis. Note that the germ band elongates along the A/P axis and shortens along the D/V axis to the same extent in the same time window (about 30 min).

In certain mutants that affect A/P patterning in the embryo, the

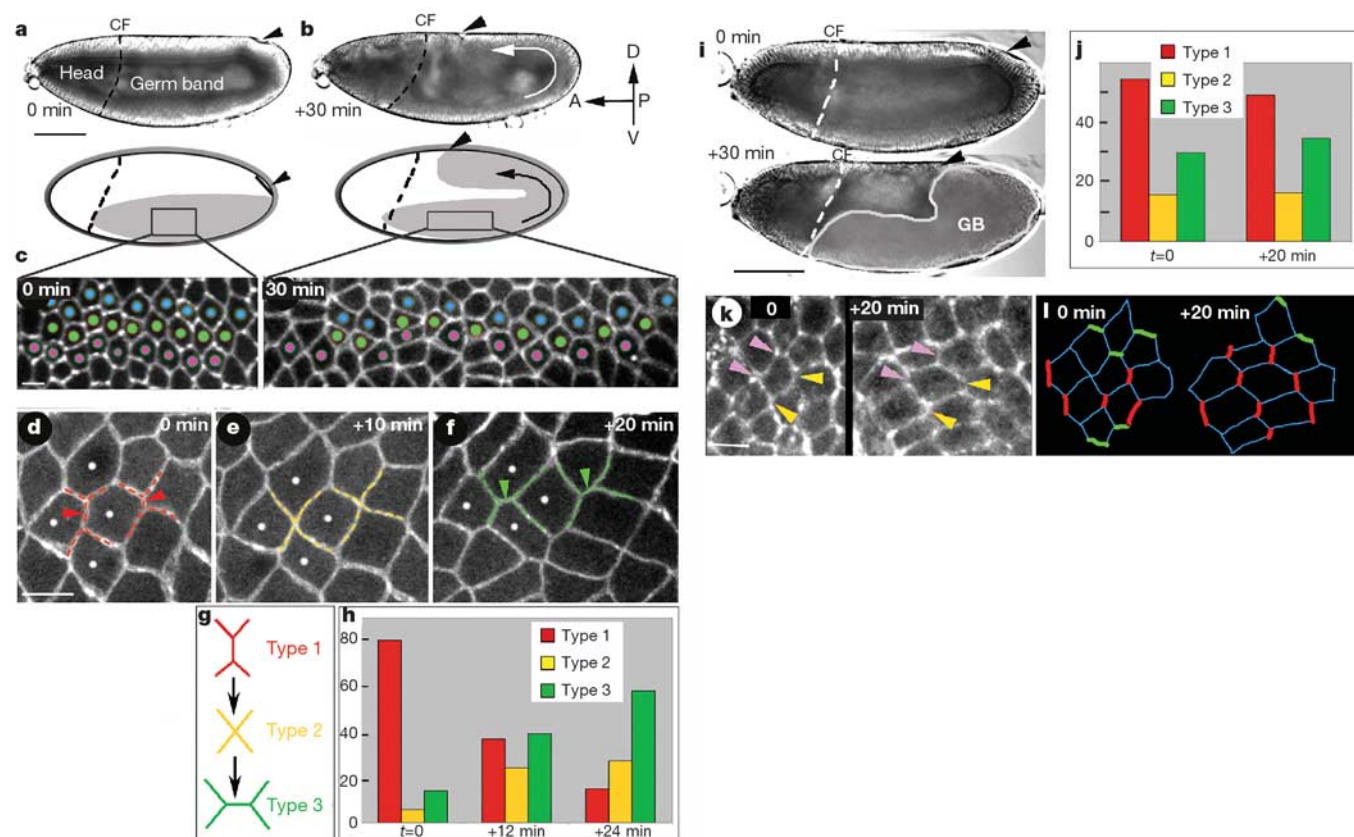


Figure 1 Polarized junction remodelling underlies planar intercalation. All the panels have the same orientation: dorsal (D) at the top, anterior (A) to the left. **a, b**, *Drosophila* embryo at the onset (**a**) of GBE and 30 min later (**b**). The germ band is posterior to the cephalic furrow (CF) and folds back dorsally (arrow) from its posterior end (arrowhead) after intercalation in the ventral–lateral region (grey). **c**, An E-cadherin–GFP fusion protein outlines the cell junctions in the boxed regions shown in **a** and **b** at corresponding time points. **d–g**, Cell contacts in the type 1 configuration (red dashed line in **d** and **g**) progress towards the type 2 (yellow dashed line in **e** and **g**) and to the type 3 configuration (green

dashed line in **f** and **g**). **h**, Percentage of type 1, type 2 and type 3 configurations at successive stages of GBE. **i**, GBE defects in a *Kr(RNAi)* mutant embryo. **j**, Distribution of type 1–type 3 configurations in *Kr(RNAi)* mutants. **k**, Junction remodelling is reduced and bi-directional in a *Kr(RNAi)* mutant. Pink arrows point to ‘backward’ transitions; orange arrows point to type 1 junctions that fail to remodel in 20 min. **l**, Schematic representation of junction remodelling with type 1 junctions (red) and type 3 junctions (green). Scale bar, 100 μ m in **a, b**; 5 μ m elsewhere.

germ band does not elongate and intercalation is affected³. If planar junction remodelling is required for axis elongation, then it should also be disrupted under these mutant conditions. The gap gene *Krüppel* (*Kr*) is required to specify distinct cell fates along the A/P axis in the central region of the germ band^{9,10}. Elimination of the activity of *Kr* by RNA-mediated interference (RNAi) causes an incomplete elongation of the germ band (Fig. 1i), a fully penetrant defect similar to that of *Kr* mutant embryos. In *Kr*(RNAi) mutant embryos, cell intercalation is strongly reduced in the central region of the epithelium. Correspondingly, type 1, type 2 and type 3 junction configurations are still observed but no change in the quantitative distribution is observed with time (Fig. 1j). This defect could occur either because cells are 'frozen' in a quasi-immobile state or because junctions are remodelled, but with no specific spatial order. Detailed analysis shows that cell intercalation is strongly reduced (junction remodelling occasionally stalls for as long as 20 min; Fig. 1k). In addition, when junction remodelling occurs, the transitions between type 1, type 2 and type 3 configurations are no longer unidirectional but tend to be more random (Fig. 1k, l). In some cases, complete 'backward' transitions are observed (Fig. 1k, pink arrows). This suggests that, in a *Kr* mutant,

GBE is reduced because junction remodelling is strongly reduced and no longer polarized in the plane of the epithelium and that, as a result, intercalation is slower and no longer oriented.

Unidirectional junction remodelling could be driven by an external force at the posterior end of the embryo, for example, associated with posterior-midgut invagination (PMI) (see Supplementary Fig. S1a), thus pulling the tissue along its axis of elongation. Alternatively, junction remodelling could depend on intrinsic forces acting locally at cell boundaries (Supplementary Fig. S1a). In the *folded-gastrulation* (*fog*) mutant, PMI and mesoderm invagination (MI) are inhibited¹¹, removing possible contributions of morphogenetic processes at the boundaries. In *fog*(RNAi) mutant embryos, PMI (Supplementary Fig. S1b) and MI (not shown) are blocked as in null mutant embryos (not shown), but we find that intercalation and junction remodelling occur normally (Supplementary Fig. S1c, e). To accommodate the space required for intercalation in the *fog* mutant, the intercalating tissue can no longer extend dorsally, because this requires PMI. Instead, intercalation causes a folding of the tissue (Supplementary Fig. S1a, c, d). This argues against the possibility that intercalation is driven by a posterior pulling force, indicating instead that it might be controlled by local mechanisms at cell junctions. The data also indicate that intercalation might be so robust that it can overcome the mechanical resistance generated during tissue compression.

The cellular effectors of these mechanisms have to fulfil several requirements. The proteins should be localized at junctions and have a localization polarized in the plane of the epithelium. Moreover, this localization should be under the control of the embryonic A/P patterning system. Finally, the proteins should be functionally required for axis elongation.

Among other constituents of apical junctions, myosin II heavy chain (encoded by the *zipper* (*zip*) gene¹²), and myosin II regulatory light chain (encoded by the *spaghetti squash* (*sqh*) gene¹³) co-localize with the β -catenin/E-cadherin complex during GBE⁴. *Sqh* and *Zip* form a very stable heterotetrameric complex that associates with F-actin and is subsequently activated by *Sqh* phosphorylation¹⁴ (Fig. 2a). We find that *Zip* localizes at apical junctions (Fig. 2d) only in intercalating cells, that is, in the ventrolateral ectoderm, and not in the non-intercalating dorsal ectoderm (Fig. 2b, c). *Zip* is not detected along the basal-lateral surface (Fig. 2d). Similar observations were made with a fully functional *Sqh*-GFP fusion protein (not shown). However, unlike β -catenin and E-cadherin, *Zip* and *Sqh*-GFP are not uniformly localized at junctions and show a polarized localization in the plane of the epithelium (Fig. 2e-g). Time-lapse analysis revealed that *Sqh*-GFP is present at all junctions but that it is enriched along shrinking type 1 junctions (Fig. 2g; $t = 0$, 85% of type 1 junctions, $n = 46$). *Sqh*-GFP is still strongly enriched at the vertex of type 2 junctions (Fig. 2g; $t = 20$ min, 87.5%, $n = 16$) but is only present at low levels along the expanding type 3 junctions (Fig. 2g; $t = 33$ min, 90%, $n = 32$). Thus, *Sqh*-GFP is enriched in shrinking cell junctions and reduced as they expand. F-actin localization revealed by phalloidin staining shows no planar polarized distribution at junctions (data not shown). Previous studies showed that *Slam*, a protein that co-localizes with myosin II and is required for membrane invagination during cellularization^{15,16}, localizes in a planar polarized fashion at junctions when ectopically expressed during GBE. In this ectopic situation, *Slam* serves as a cell marker of planar cell intercalation. Like myosin II, *Slam*'s planar polarized localization reflects a specific enrichment at type 1 junctions (Fig. 2i); ectopic *Slam* is absent from the new type 3 junctions (Fig. 2j, arrowheads). The planar polarized localization of myosin II is strongly affected in *Kr*(RNAi) mutants in which junction remodelling is strongly reduced and no longer polarized. Instead of being enriched at type 1 junctions, *Sqh*-GFP is globally detected at lower levels and present in dispersed speckles at tricellular junctions (Fig. 2h; 9% of enriched *Sqh*-GFP at type 1 junctions, $n = 56$, compared with 85% in the wild type, $n = 46$).

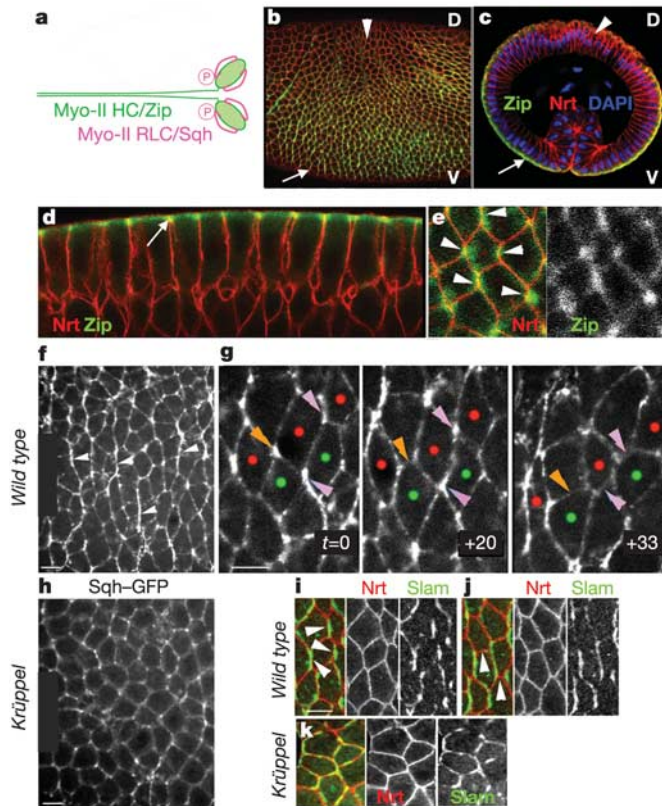


Figure 2 Myosin II-polarized localization. **a**, Myosin II regulatory light chain (Myo-II-RLC)/Sqh and myosin II heavy chain (Myo-II-HC)/Zip hetero-tetramer. **b–e**, Double staining for Zip (green) and Nrt (red) during GBE shown from the surface (**b**) and in cross-section (**c**). Zip is present at cell junctions in the ventral (V) lateral intercalating ectoderm (**b, c**, arrow, **d**) and cytoplasmic in the non-intercalating dorsal (D) ectoderm (**b, c**, arrowheads). Zip is specifically enriched in type-1 junctions (**e**). **f**, Enrichment of Sqh-GFP at type 1 junctions (arrowheads). **g**, Time-lapse sequence of Sqh-GFP (times in minutes), showing the enrichment at type 1 junctions (**g**, $t = 0$, pink and orange arrowheads), in type 2 junctions (+20 min, pink arrowheads) and the reduced localization at type 3 junctions (+33 min, pink and orange arrowheads). **h**, In a *Kr*(RNAi) mutant embryo, Sqh-GFP is present at lower levels with severe planar polarity defects (9% of type 1 junctions enriched for Sqh-GFP instead of 85% in the wild type). **i–k**, Ectopic *Slam* localization at junctions (green) also concentrates in type 1 junctions (**i**, arrowheads), but is absent from type-3 junctions (**j**, arrowheads). **k**, In a *Kr*(RNAi) mutant, ectopic *Slam* is also localized more uniformly. Scale bar, 5 μ m.

This loss of polarity within the plane of the epithelium is also seen with ectopic *Slam* at junctions (Fig. 2k).

Because the localization of myosin II remarkably mirrors the planar polarized remodelling of junctions in the epithelium, myosin II is a very good candidate for driving this process during GBE. Given the known role of the myosin II motor in other cellular contexts^{17–20}, a myosin-dependent polarized contractile force might direct cell intercalation by forcing E-cadherin contact remodelling through the underlying actin–myosin cytoskeleton. Myosin II is maternally provided in the early embryo. However, removing the maternal contribution is not possible because the oocyte does not develop properly when it is homozygous mutant for *zip* or *sqh*^{21,22}. In contrast, removing the zygotic contribution of *zip* and *sqh* has been reported to cause late developmental defects¹⁴. However, using time-lapse analysis of a large number of embryos mutant for different alleles of *zip* and *sqh*, we also found a clear requirement for myosin II during GBE, with GBE defects in up to 20% of embryos from heterozygous *zipper* or *sqh* parents (Fig. 3a–c). We then looked at cell intercalation by using Ecad–GFP in a *zip*² mutant and found that indeed intercalation is severely disrupted, leading to a quasi-‘frozen’ tissue (Fig. 3d, e). Junction remodelling did not occur, and type 1 junctions could remain unchanged for at least 40 min (Fig. 3f–h). In spite of this, the epithelium remained intact and in particular E-cadherin localization was normal at the junctions. Rho kinase (Rok) activates the Zip/Sqh complex by the phosphorylation of Sqh²³. We used a specific inhibitor of Rok (Y-27632 (refs 24, 25)) to further address the role of myosin II during planar junction remodelling. We injected Y-27632 during cellularization and found that Sqh–GFP recruitment at the junctions was strongly reduced (Fig. 4a, b), although cell junctions were

normal (Fig. 4d–h). GBE was severely affected in a dose-dependent fashion (Fig. 4c) and this phenotype was rescued when the injection was performed in a strain heterozygous for a transgene that expresses low levels of a phospho-mimetic *sqh* mutant (*sqhE20E21*) immune to Rok inhibition^{23,26} (Fig. 4c), which by itself did not cause any phenotype (not shown). Both cell intercalation (Fig. 4d, e) and planar junction remodelling (Fig. 4f–h) were severely affected after Y-27632 injection, as in the *zip* mutant embryos (Fig. 3h).

The data presented here identify a particularly simple and novel cellular mechanism for planar intercalation within an epithelium that undergoes elongation (Fig. 4i, j). Planar epithelial intercalation

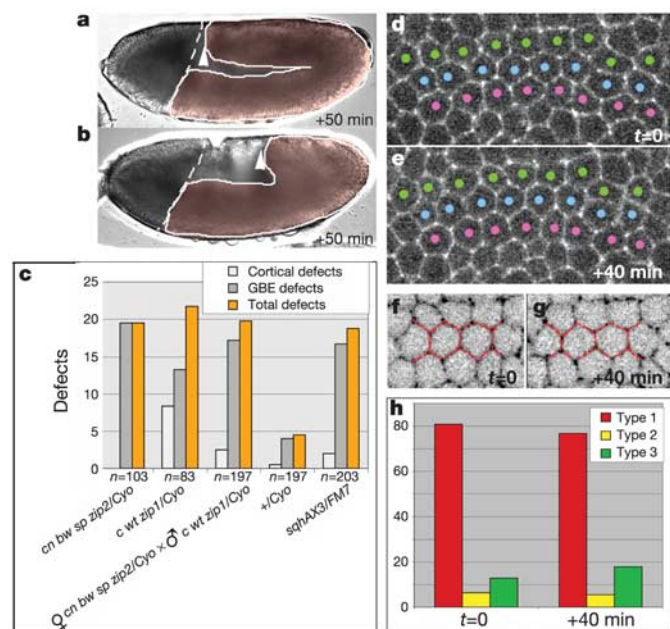


Figure 3 Myosin II is required for junction remodelling. **a, b**, GBE defect in an embryo mutant for *zip*² (**b**) in comparison with a wild-type embryo (**a**). The germ band is shaded in red and its posterior end is marked with a white arrowhead. **c**, Statistics of GBE defects in various *zipper* and *sqh* mutant conditions. All embryos were revealed by time-lapse microscopy and GBE defects are similar to or stronger than the embryo shown in **b**. A few embryos had some defects at the cortex of the embryos before cellularization (white bars). We scored only embryos with GBE defects that had cellularized normally (grey bars). **d, e**, Cells expressing E-cadherin–GFP in a *zip*² mutant embryo are marked with a coloured dot and tracked from the onset of GBE (**d**) until 40 min later (**e**). Cell intercalation is blocked in this embryo. **f, g**, Junction remodelling is ‘frozen’ from the onset of GBE (**f**) until 40 min later (**g**). **h**, The percentage of type-1, type-2 and type-3 junctions is unchanged at 40-min intervals.

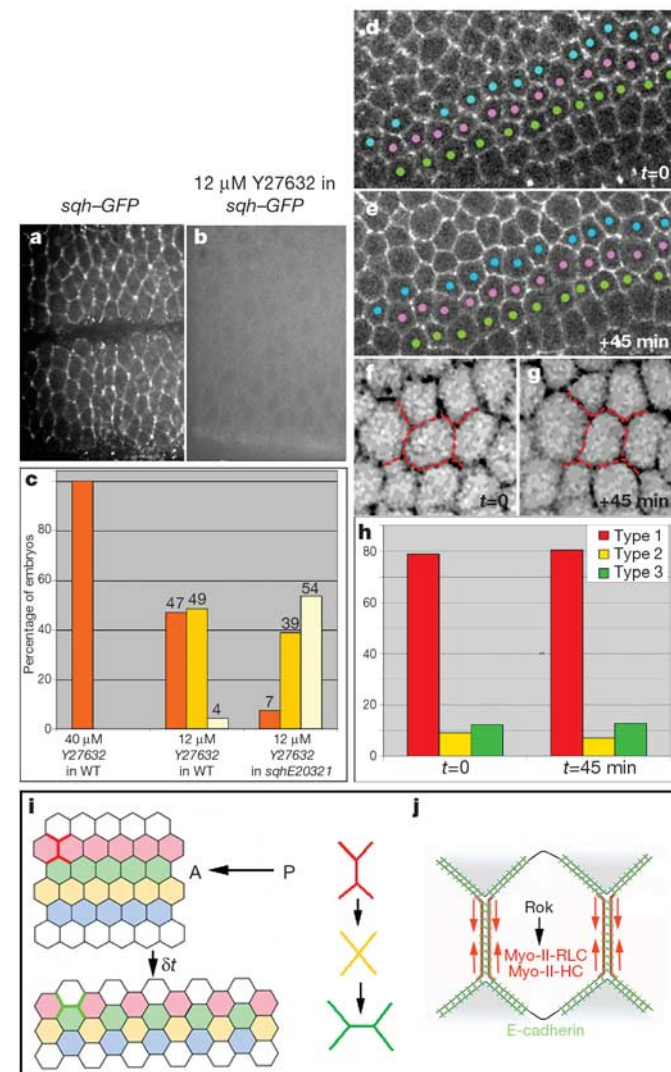


Figure 4 Role of Rok during intercalation. **a, b**, Sqh–GFP localization at cell junctions in a control embryo (**a**) and an embryo injected with the Rok inhibitor Y-27632 (**b**). **c**, GBE defects scored after Y-27632 injection in control (wild-type; WT) embryos (left and middle) and embryos expressing SqhE20E21 (right). Embryos with the most severe GBE defects (less than 35%) are shown in orange, intermediate elongation defects (35–70%) in yellow and near wild-type GBE (more than 70%) in pale yellow. **d, e**, Cells expressing E-cadherin–GFP in an embryo injected with Y-27632 are marked with a coloured dot and tracked from the onset of GBE (**d**) until 45 min later (**e**). Note that cell intercalation is blocked. This embryo elongates its germ band less than 35%. **f, g**, Junction remodelling is also ‘frozen’ from the onset of GBE (**f**) until 45 min later (**g**). **h**, The percentage of type-1, type-2 and type-3 junctions is unchanged at 45-min intervals. **i, j**, Model: planar cell intercalation depends on the polarized remodelling of junctions along the A/P axis (**i**). This irreversible process is controlled by the polarized recruitment of myosin II (Myo-II) in a subset of cell junctions (**j**), thereby forcing polarized junction remodelling.

depends on the polarized remodelling of cell junctions. Junctions are remodelled through the polarized recruitment of myosin II within the plane of the epithelium. The contractile activity of myosin II might create a local tension that orients the disassembly of E-cadherin junctions. Alternatively, myosin II might interact with and control the activity of regulators of E-cadherin stabilization. Myosin II is required to progress from the type 1 to the type 2 configuration, counteracting the effect of junction formation and stabilization present at all cell junctions in epithelial cells before and during GBE. In principle, when cell–cell contacts are in the type 2 configuration, new contacts would form either by reverting to the type 1 configuration or by progressing to the type 3 configuration. However, we propose that the polarized localization of myosin II counteracts attempts to form new type 1 junctions, thus allowing the formation of only type 3 junctions. In other words, the polarized localization of myosin II determines the irreversible transformation from type 1 to type 3. This basic cellular mechanism could be involved in a variety of biological contexts in different organisms. Epithelial tubes, for instance, elongate during organogenesis from *Drosophila* to vertebrates^{2,27,28}. Nevertheless, the mechanisms that orient the polarized recruitment of myosin might vary. During GBE, local signalling between adjacent cells along the A/P axis might suffice to recruit myosin II along A/P junctions specifically. In other contexts, long-range signalling by the Frizzled/Dishevelled planar cell polarity pathway might be involved^{1,29}. The work presented focuses on the cellular effectors of such signalling pathways and shows how they might underlie a variety of morphogenetic processes. □

Methods

Imaging techniques

Confocal time-lapse images were collected using a spinning-disc confocal head (Perkin Elmer) run by Metamorph software on an inverted microscope (Zeiss). Phase-contrast time-lapse images were collected on an inverted microscope (Zeiss) and a programmable motorized stage to record different positions over time (Mark&Find module from Zeiss). The system was run with AxioVision software (Zeiss) and allowed the acquisition of large time-lapse data sets in mutant or injected embryos.

RNA interference

dsRNA probes to *Krüppel* (between positions 494 and 1268 of the coding sequence) and *fog* (860 bases, –24 from 5' of the ATG to position 837 3' of the ATG) were synthesized *in vitro* and injected in early embryos (less than 1 h old) at 0.5 µg µl⁻¹.

Injectations

Y-27632 was prepared as a 50 mM stock in water and injected at 600 µM at the onset of cellularization. The dilution factor of Y-27632 in the syncytial embryo was estimated to be 1:50, so we estimate the local concentration as about 12 µM.

Genetics

The following stocks were used. *ubi-E-cad-GFP* homozygous flies (chromosome II) express a functional E-cadherin–GFP fusion⁸. The functional *Sqh-GFP* fusion was expressed with a genomic rescue transgene in a *sqhAX3*-null mutant background²⁶. The following stocks were used to assess the role of myosin II in Fig. 3: *cn bw sp. zip²/Cyo*, *c wt zip¹/Cyo*, *EcadGFP sp. zip²/Cyo* and *y w sqhAX3/FM7*. The ectopic expression of a *UAS-SlamHA* transgene during GBE was monitored in embryos laid by *matexub-Gal4VP16* mothers crossed to *UAS-slamHA* males.

Received 27 February; accepted 23 April 2004; doi:10.1038/nature02590.

1. Wallingford, J. B., Fraser, S. E. & Harland, R. M. Convergent extension: the molecular control of polarized cell movement during embryonic development. *Dev. Cell* **2**, 695–706 (2002).
2. Keller, R. Shaping the vertebrate body plan by polarized embryonic cell movements. *Science* **298**, 1950–1954 (2002).
3. Irvine, K. D. & Wieschaus, E. Cell intercalation during *Drosophila* germband extension and its regulation by pair-rule segmentation genes. *Development* **120**, 827–841 (1994).
4. Lecuit, T. & Wieschaus, E. Junctions as organizing centers in epithelial cells? A fly perspective. *Trends Cell Biol.* **10**, 92–97 (2002).
5. Knust, E. & Bossinger, O. Composition and formation of intercellular junctions in epithelial cells. *Science* **298**, 1955–1959 (2002).
6. Nelson, W. J. Adaptation of core mechanisms to generate cell polarity. *Nature* **422**, 766–774 (2003).
7. Adams, C. L. & Nelson, W. J. Cytomechanics of cadherin-mediated cell–cell adhesion. *Curr. Opin. Cell Biol.* **10**, 572–577 (1998).
8. Oda, H. & Tsukita, S. Real-time imaging of cell–cell adherens junctions reveals that *Drosophila* mesoderm invagination begins with two phases of apical constriction of cells. *J. Cell Sci.* **114**, 493–501 (2001).
9. Rivera-Pomar, R. & Jackle, H. From gradients to stripes in *Drosophila* embryogenesis: filling in the gaps. *Trends Genet.* **12**, 478–483 (1996).
10. Small, S., Kraut, R., Hoey, T., Warrior, R. & Levine, M. Transcriptional regulation of a pair-rule stripe

in *Drosophila*. *Genes Dev.* **5**, 827–839 (1991).

11. Costa, M., Wilson, E. T. & Wieschaus, E. A putative cell signal encoded by the folded gastrulation gene coordinates cell shape changes during *Drosophila* gastrulation. *Cell* **76**, 1075–1089 (1994).
12. Kiehart, D. P. Molecular genetic dissection of myosin heavy chain function. *Cell* **60**, 347–350 (1990).
13. Karsenti, R. E. *et al.* The regulatory light chain of nonmuscle myosin is encoded by spaghetti-squash, a gene required for cytokinesis in *Drosophila*. *Cell* **65**, 1177–1189 (1991).
14. Young, P. E., Richman, A. M., Ketchum, A. S. & Kiehart, D. P. Morphogenesis in *Drosophila* requires nonmuscle myosin heavy chain function. *Genes Dev.* **7**, 29–41 (1993).
15. Lecuit, T., Samanta, R. & Wieschaus, E. *slam* encodes a developmental regulator of polarized membrane growth during cleavage of the *Drosophila* embryo. *Dev. Cell* **2**, 425–436 (2002).
16. Stein, J. A., Broihier, H. T., Moore, L. A. & Lehmann, R. Slow as Molasses is required for polarized membrane growth and germ cell migration in *Drosophila*. *Development* **129**, 3925–3934 (2002).
17. Kiehart, D. P., Galbraith, C. G., Edwards, K. A., Rickoll, W. L. & Montague, R. A. Multiple forces contribute to cell sheet morphogenesis for dorsal closure in *Drosophila*. *J. Cell Biol.* **149**, 471–490 (2000).
18. Hutson, M. S. *et al.* Forces for morphogenesis investigated with laser microsurgery and quantitative modeling. *Science* **300**, 145–149 (2003).
19. Wolf, W. A., Chew, T. L. & Chisholm, R. L. Regulation of cytokinesis. *Cell. Mol. Life Sci.* **55**, 108–120 (1999).
20. Glotzer, M. Animal cell cytokinesis. *Annu. Rev. Cell Dev. Biol.* **17**, 351–386 (2001).
21. Jordan, P. & Karsenti, R. Myosin light chain-activating phosphorylation sites are required for oogenesis in *Drosophila*. *J. Cell Biol.* **139**, 1805–1819 (1997).
22. Wheatley, S., Kulkarni, S. & Karsenti, R. *Drosophila* nonmuscle myosin II is required for rapid cytoplasmic transport during oogenesis and for axial nuclear migration in early embryos. *Development* **121**, 1937–1946 (1995).
23. Winter, C. G. *et al.* *Drosophila* Rho-associated kinase (Drok) links Frizzled-mediated planar cell polarity signaling to the actin cytoskeleton. *Cell* **105**, 81–91 (2001).
24. Uehata, M. *et al.* Calcium sensitization of smooth muscle mediated by a Rho-associated protein kinase in hypertension. *Nature* **389**, 990–994 (1997).
25. Narumiya, S., Ishizaki, T. & Uehata, M. Use and properties of ROCK-specific inhibitor Y-27632. *Methods Enzymol.* **325**, 273–284 (2000).
26. Royou, A., Sullivan, W. & Karsenti, R. Cortical recruitment of nonmuscle myosin II in early syncytial *Drosophila* embryos: its role in nuclear axial expansion and its regulation by Cdc2 activity. *J. Cell Biol.* **158**, 127–137 (2002).
27. Johansen, K. A., Iwaki, D. D. & Lengyel, J. A. Localized JAK/STAT signaling is required for oriented cell rearrangement in a tubular epithelium. *Development* **130**, 135–145 (2003).
28. Jazwinska, A., Ribeiro, C. & Affolter, M. Epithelial tube morphogenesis during *Drosophila* tracheal development requires Piopio, a luminal ZP protein. *Nature Cell Biol.* **5**, 895–901 (2003).
29. Tree, D. R., Ma, D. & Axelrod, J. D. A three-tiered mechanism for regulation of planar cell polarity. *Semin. Cell Dev. Biol.* **13**, 217–224 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. M. Philippe and J. Della Rovere for their technical help; A. Pelissier for the staining shown in Fig. 2c; R. Karsenti, D. Kiehart and H. Oda for sending key reagents; F. Pilot and C. Henderson for insights; F. Schweisguth for a useful suggestion; and S. Cohen, J. Ewbank, C. Henderson, S. Kerridge, F. Pilot, O. Pourquie and J. Pradel for comments on the manuscript. Research in the Lecuit laboratory is supported by an ATIP grant from the CNRS, by the Fondation pour la Recherche Médicale, the Association pour la Recherche contre le Cancer and the EMBO Young Investigator Programme.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.L. (lecuit@ibdm.univ-mrs.fr).

Structure of a complex between a voltage-gated calcium channel β -subunit and an α -subunit domain

Filip Van Petegem, Kimberly A. Clark, Franck C. Chatelain & Daniel L. Minor Jr

Cardiovascular Research Institute, Departments of Biochemistry and Biophysics, and Cellular and Molecular Pharmacology, University of California San Francisco, 513 Parnassus Avenue, Box 0130, San Francisco, California 94143, USA

Voltage-gated calcium channels (Ca_vs) govern muscle contraction, hormone and neurotransmitter release, neuronal migration, activation of calcium-dependent signalling cascades, and synaptic input integration¹. An essential Ca_v intracellular protein, the β -subunit (Ca_v β)^{1,2}, binds a conserved domain (the α -interaction domain, AID) between transmembrane domains I

and II of the pore-forming α_1 subunit³ and profoundly affects multiple channel properties such as voltage-dependent activation², inactivation rates², G-protein modulation⁴, drug sensitivity⁵ and cell surface expression^{6,7}. Here, we report the high-resolution crystal structures of the $\text{Ca}_v\beta_{2a}$ conserved core, alone and in complex with the AID. Previous work suggested that a conserved region, the β -interaction domain (BID), formed the AID-binding site⁸; however, this region is largely buried in the $\text{Ca}_v\beta$ core and is unavailable for protein–protein interactions. The structure of the AID– $\text{Ca}_v\beta_{2a}$ complex shows instead that $\text{Ca}_v\beta_{2a}$ engages the AID through an extensive, conserved hydrophobic cleft (named the α -binding pocket, ABP). The ABP–AID interaction positions one end of the $\text{Ca}_v\beta$ near the intracellular end of a pore-lining segment, called IS6, that has a critical role in Ca_v inactivation^{9,10}. Together, these data suggest that $\text{Ca}_v\beta$ s influence Ca_v gating by direct modulation of IS6 movement within the channel pore.

The 1.97 Å resolution structure of the $\text{Ca}_v\beta_{2a}$ core shows that $\text{Ca}_v\beta$ s comprise two well-conserved domains (Fig. 1a). The first, an SH3 fold, contains five antiparallel β -strands ($\beta 1$ – $\beta 5$), a 3_{10} helix ($\eta 1$), and two α -helices ($\alpha 1$ and $\alpha 2$) that lie amino-terminal to $\beta 1$ and carboxy-terminal to $\beta 4$, respectively. The strand that completes the SH3 fold, $\beta 5$ (residues 217–224), is separated in the primary structure from the core of the SH3 domain by approximately 70 residues (variable domain 2, V2, a site of splice variation and amino acid insertions and deletions²) that are absent from the structure (Fig. 1b). The second conserved domain consists of a five-stranded parallel β -sheet ($\beta 6$ – $\beta 10$), surrounded by six α -helices ($\alpha 3$ – $\alpha 8$) and two 3_{10} helices ($\eta 2$ and $\eta 3$), and is related to the core of nucleotide kinase enzymes.

$\text{Ca}_v\beta$ s share structural features with membrane-associated guanylate kinases (MAGUKs), a protein scaffold family that organizes signalling components near membranes¹¹. MAGUKs contain one or more PDZ domains N-terminal to an SH3 domain, a bridging region known as a HOOK domain and a nucleotide kinase domain^{11,12}. PDZ domains are approximately 100 residues long. The $\text{Ca}_v\beta_{2a}$ structure indicates that $\text{Ca}_v\beta$ s lack N-terminal PDZ domains. There are too few residues N-terminal to the SH3 domain (even in $\text{Ca}_v\beta_{1b}$, the $\text{Ca}_v\beta$ with the longest (55 amino acids) N-terminal variable region 1, V1) to fold as a PDZ domain. The absence of a PDZ domain distinguishes $\text{Ca}_v\beta$ s from canonical MAGUK proteins.

Comparison of $\text{Ca}_v\beta_{2a}$ with a representative MAGUK, PSD-95 (refs 12, 13), reveals other differences. Superposition of the nucleotide kinase domains shows that the relative orientations of the SH3 and nucleotide kinase domains differ by approximately 90°, an arrangement that makes $\text{Ca}_v\beta_{2a}$ a more elongated structure (Fig. 2a). The nucleotide kinase domain of MAGUKs is homologous to guanylate kinases and retains guanosine monophosphate (GMP) binding, but key residues for enzymatic function are missing¹². The four-stranded β -sheet nucleotide kinase subdomain that binds GMP in MAGUKs is absent in $\text{Ca}_v\beta_{2a}$ (Fig. 2a). Furthermore, two $\text{Ca}_v\beta_{2a}$ loops (between $\beta 7$ – $\eta 2$ and $\beta 8$ – $\beta 9$) occlude part of the binding site for the GMP guanosine ring. Thus, the $\text{Ca}_v\beta$ nucleotide kinase domain seems to have lost the ability to bind nucleotides.

The structures of $\text{Ca}_v\beta_{2a}$ and PSD-95 SH3 domains are similar (Fig. 2b). Neither is compatible with canonical modes of proline-rich ligand binding. Both lack the aromatic residues necessary for ligand engagement¹³, and the surface that would bind polyproline ligands is blocked by the $\alpha 2$ helix^{12,13}. In PSD-95, residues C-terminal to the nucleotide kinase domain contribute an extra SH3 β -strand that is absent from canonical SH3 domains¹³ and absent in $\text{Ca}_v\beta_{2a}$. The HOOK domain, present in MAGUKs and $\text{Ca}_v\beta_{2a}$, bridges SH3 β -strands $\beta 4$ and $\beta 5$ and comprises $\alpha 2$ and variable domain 2 of $\text{Ca}_v\beta_{2a}$. HOOK domains are important regulatory regions for interactions of MAGUKs with other proteins^{11,12,14} and may serve a similar function for $\text{Ca}_v\beta$ protein–protein interactions.

$\text{Ca}_v\beta$ s exert their effects on Ca_v function by binding the pore-forming α_1 subunit at a conserved, 18-residue sequence located between membrane domains I and II (the AID)^{3,15} (Fig. 3a). Interpretation of mutagenesis and biochemical studies suggested that $\text{Ca}_v\beta$ –AID interactions occur through a 41-residue segment ($\text{Ca}_v\beta_{2a}$ residues 212–252) termed the β -interaction domain (BID)^{8,16}. The $\text{Ca}_v\beta_{2a}$ structure shows that the central region of the BID, which includes the residues previously thought to be important for BID–AID interactions⁸, is entirely buried and cannot participate directly in protein–protein interactions (Fig. 3b). Two putative BID phosphorylation sites^{8,16} are also buried in this region. Given the extent of burial, mutations used to determine the relative importance of residues involved in BID–AID interactions (for example, proline to arginine)⁸ are likely to have abolished AID binding by disrupting the folded structure of the nucleotide kinase domain rather than by perturbing direct AID contacts. Although it would appear that the $\text{Ca}_v\beta_{2a}$ structure conflicts with previous data,

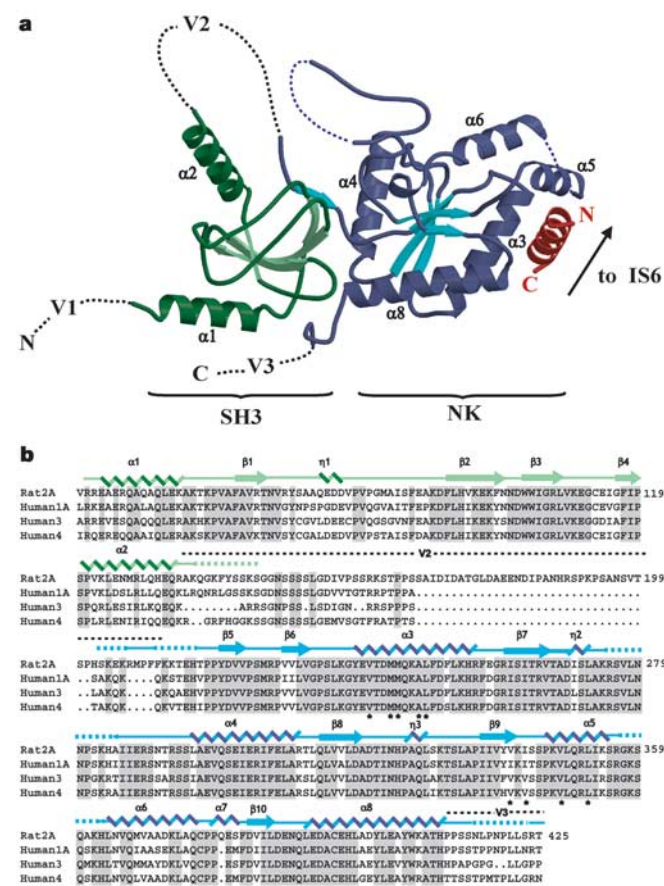


Figure 1 Structure of the $\text{Ca}_v\beta_{2a}$ – $\text{Ca}_v1.2$ AID complex. **a**, Ribbon diagram of the complex. Dashed lines indicate regions absent from the structures. SH3 and nucleotide kinase (NK) domains are shown in green and blue, respectively. The AID is shown in red. $\text{Ca}_v\beta_{2a}$ α -helices are labelled. Variable regions V1, V2 and V3 are indicated. The $\text{Ca}_v\beta_{2a}$ unbound structure is similar to that shown here for the complex. The arrow indicates where the AID connects to transmembrane segment IS6. **b**, Sequence alignment of representatives of each $\text{Ca}_v\beta$ isoform. The top sequence shows residues 40–425 of rat $\text{Ca}_v\beta_{2a}$. Numbers on the right denote each line's terminal residue. Shading denotes residues identical among isoforms. The two $\text{Ca}_v\beta_{2a}$ domains used for crystallization are indicated in green and blue, respectively. Secondary structure elements are indicated: α , α -helix; η , 3_{10} helix; β , β -strand. Dashed lines indicate residues present in the crystallized constructs but absent in the electron density. Location of the V2 and part of the V3 regions are shown. Asterisks identify residues that contribute side-chain contacts to the AID-binding pocket; diamonds mark side chains with direct hydrogen bonds to the AID.

most of the supporting evidence for the BID–AID interaction relied on indirect functional experiments and direct BID–AID binding was never demonstrated^{8,16}.

If the BID does not interact directly with the AID, how do $\text{Ca}_v\beta_{2a}$ and the AID interact? To answer this, we solved the high-resolution (2.00 Å) structure of a complex between the conserved $\text{Ca}_v\beta_{2a}$ core and an 18-residue peptide containing the AID from the L-type channel $\text{Ca}_v1.2$. The electron density reveals the first 16 residues of the AID and the location of the binding pocket on $\text{Ca}_v\beta_{2a}$ (Fig. 3c). Overall, the $\text{Ca}_v\beta_{2a}$ structure is very similar to the unbound form (root-mean-square deviation (RMSD)_{C α} = 0.397 Å) and bears only a few conformational changes in side chains near the AID-binding pocket ($\text{Ca}_v\beta_{2a}$ residues M244, N390, E393 and R351). The AID forms an α -helix that is anchored to the binding pocket through a set of conserved residues (AID residues L434, G436, Y437, W440 and I441) that are important for $\text{Ca}_v\beta$ binding and for α -subunit modulation by $\text{Ca}_v\beta$ s^{8,15,17}. These residues bind a deep groove that we call the α -binding pocket (ABP), formed by helices $\alpha 3$, $\alpha 5$ and

$\alpha 8$ of the $\text{Ca}_v\beta_{2a}$ nucleotide kinase domain at a site distal to the SH3 domain (Figs 1a, 3b and 4a).

The complex buries approximately 730 Å² of ABP surface, of which about 350 Å² are hydrophobic. AID side chains D433, W440 and Q443 make direct hydrogen bonds to the ABP (Supplementary Fig. 1). The depth and extent of burial of the aromatic AID positions Y437 and W440 is particularly striking (Fig. 4b, c). The AID Y437 hydroxyl group is central to a buried hydrogen bond network comprising three water molecules, AID residue D433 and five $\text{Ca}_v\beta_{2a}$ residues (Supplementary Fig. 1). Mutation of this tyrosine to phenylalanine greatly diminishes AID– $\text{Ca}_v\beta$ binding¹⁵. The extensive AID–ABP interactions are consistent with biochemical experiments demonstrating strong AID– $\text{Ca}_v\beta$ interaction (dissociation constant ~6–20 nM)¹⁸. The $\text{Ca}_v\beta$ side chains that contact the AID are highly conserved among $\text{Ca}_v\beta$ isoforms (see Figs 1b, 3a and 4a; see also Supplementary Fig. 1). Thus, both binding partners engage each other through conserved residues to create the AID–ABP interaction.

Interactions between the $\text{Ca}_v\alpha_1$ and $\text{Ca}_v\beta$ subunits markedly

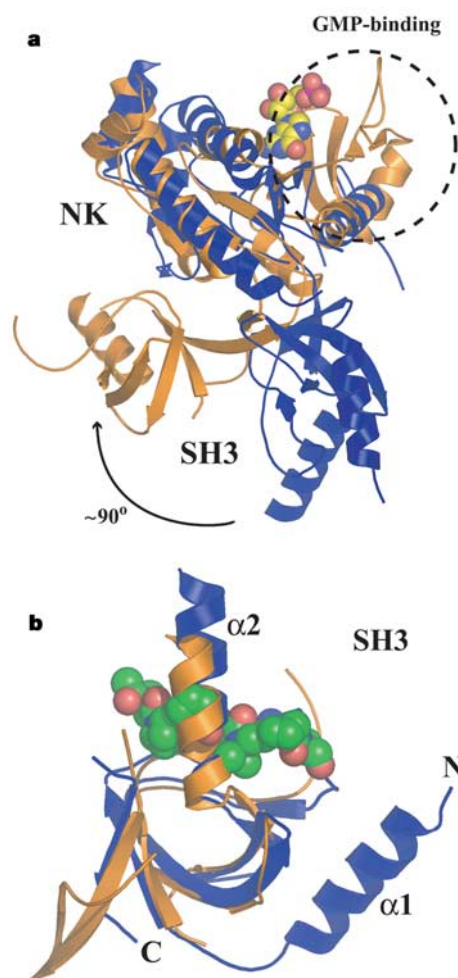


Figure 2 Structural comparisons between PSD-95 (gold) and $\text{Ca}_v\beta_{2a}$ (blue). **a**, Superposition of $\text{Ca}_v\beta_{2a}$ and PSD-95 nucleotide kinase domains (RMSD_{C α} = 3.9 Å). The dashed circle indicates the guanosine-monophosphate (GMP)-binding domain present in PSD-95 but absent in $\text{Ca}_v\beta_{2a}$. The guanosine monophosphate molecule bound to PSD-95 is displayed in space-filling representation. Nucleotide kinase (NK) and SH3 domains are indicated. The relative change in SH3 domain orientation is indicated. **b**, Superposition of PSD-95 and $\text{Ca}_v\beta_{2a}$ SH3 domains (RMSD_{C α} = 1.6 Å). Position of the polyproline ligand from a superposition with the Sem5 SH3 domain (Protein Data Bank code 2SEM) (RMSD_{C α} = 1.8 Å) is shown in space-filling representation. The Sem5 SH3 is not shown. The DALI server generated the superpositions (<http://www.ebi.ac.uk/dali/>).

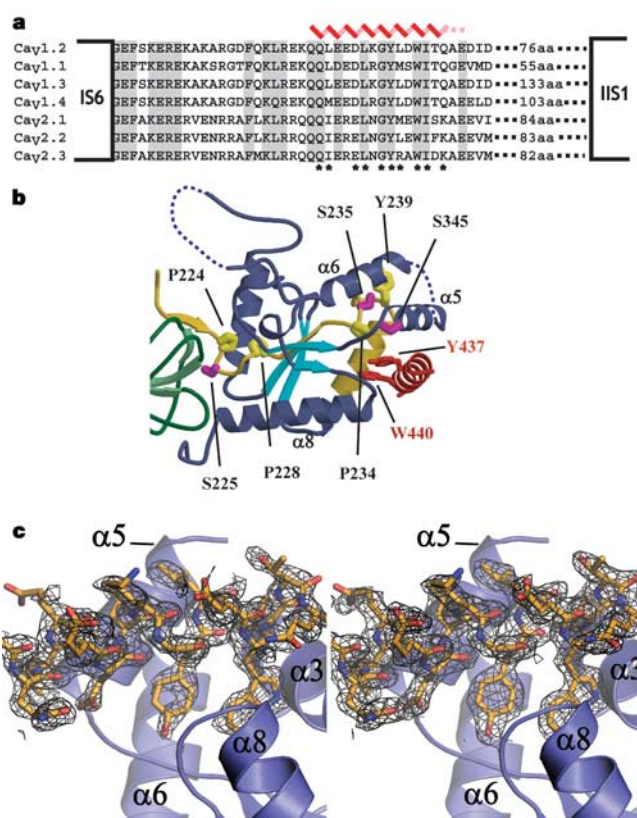


Figure 3 Features of the AID– $\text{Ca}_v\beta_{2a}$ interaction and location of the previously described BID. **a**, Sequence alignment of AID domains ($\text{Ca}_v1.2$ residues 428–445) and neighbouring residues. The positions of the last transmembrane segment of transmembrane domain I (IS6) and the first transmembrane segment of transmembrane domain II (IIS1) are shown. Secondary structure of the AID from the co-crystal structure is indicated (red). Dashed lines indicate residues absent from the electron density. Asterisks identify side-chain contacts with $\text{Ca}_v\beta_{2a}$ closer than 4 Å. **b**, Position of the previously described BID (residues 212–252; yellow)^{3,8}. Residues previously proposed to mediate AID–BID interactions (P224, P228, P234, Y239) are indicated and have relative accessibilities of 1.4%, 0%, 0% and 32.4%, respectively. Putative PKC sites S225, S235 and S345 are also shown (magenta) and have relative accessibilities of 8.8%, 0% and 35.4%, respectively. S345 accessibility reduces to 12% in the complex. Accessibility values are relative to a tripeptide, Gly-X-Gly. **c**, The left panel shows $F_o - F_c$ electron density, contoured at 2σ , for the AID– $\text{Ca}_v\beta_{2a}$ complex before building the AID. The right panel shows final $2F_o - F_c$ density, contoured at 1σ , for the AID from the refined AID– $\text{Ca}_v\beta_{2a}$ structure (right). In both panels the final AID model is shown.

influence the cell surface expression of functional channels^{6,7}. Control of Ca_v trafficking by regulating $\text{Ca}_v\alpha_1$ – $\text{Ca}_v\beta$ interactions is emerging as an important means of modulating cellular excitability⁷. Ca_v channel subtypes are major clinical targets for drugs that treat cardiovascular disease, migraine and pain¹⁹. Development of compounds that could interfere with the AID–ABP binding interactions might provide new ways to modulate Ca_v function in pathological states.

The $\text{Ca}_v\beta$ –AID structure provides a starting point for understanding how $\text{Ca}_v\beta$ modulates numerous channel properties. G-protein $\beta\gamma$ subunits ($\text{G}\beta\gamma$) inhibit Ca_v function¹⁴; however, the sites of $\text{G}\beta\gamma$ – Ca_v interactions and precise inhibitory mechanisms remain highly controversial⁴. Biochemical experiments show that $\text{G}\beta\gamma$ binds the Ca_v2 AID and that mutations at AID positions Q1, Q2, R5, L7, G9 and Y10 (corresponding to $\text{Ca}_v1.2$ AID residues 428, 429, 432, 434, 436 and 437) abolish $\text{G}\beta\gamma$ –AID binding²⁰. In contrast, other studies suggest that the *in vitro* $\text{G}\beta\gamma$ –AID interaction is functionally irrelevant and that the relevant $\text{G}\beta\gamma$ –binding determinants lie elsewhere in the channel cytoplasmic domains^{21,22}. The structure of the $\text{Ca}_v\beta_{2a}$ –AID complex shows that three of the putative $\text{G}\beta\gamma$ –AID interacting positions (L7, G9 and Y10), which are invariant in Ca_v1 and Ca_v2 channels (Fig. 3a), are deeply buried by interactions with $\text{Ca}_v\beta$ (Fig. 4a, b; see also Supplementary Fig. 2 and Table 2). The extent of burial of these residues, which are critical for maintaining $\text{Ca}_v\beta$ –AID association^{3,15}, suggests that $\text{G}\beta\gamma$ and $\text{Ca}_v\beta$ cannot bind to the AID simultaneously. Taken together with the observation that the $\text{G}\beta\gamma$ –AID affinity is at least 10–20-fold weaker than the $\text{Ca}_v\beta$ –AID²⁰ affinity, the structure also indicates

that it is unlikely that $\text{G}\beta\gamma$ could effectively compete with $\text{Ca}_v\beta$ for AID binding without drastic structural rearrangement of the $\text{Ca}_v\beta$ –AID complex. Thus, our data lend support to the view that the major $\text{G}\beta\gamma$ interaction sites lie in other Ca_v cytoplasmic domains^{21–23}.

How might $\text{Ca}_v\beta$ s affect channel gating? Although the detailed mechanisms of Ca_v inactivation processes remain unknown, functional experiments show that the IS6 pore-lining segment has a key role¹⁰. Motions of pore-lining transmembrane segments are a common theme emerging in ion channel gating^{24,25}. Twenty-two residues separate the AID helix N terminus from the cytoplasmic end of IS6. We do not know the structure, but sequence evaluation suggests that these residues have a high helix propensity and could readily form a continuous helix between the AID helix and the presumed transmembrane helix of IS6 (Fig. 5a). AID residue 5 (E432, here) slows inactivation when negatively charged (as in Ca_v1 channels) and speeds inactivation when positively charged (as in Ca_v2 channels)^{26,27}. This residue is exposed on the surface of the AID helix (Fig. 4a) where it would be available to interact with other parts of the channel and could influence the rates of movement of the $\text{Ca}_v\beta$ –AID complex and therefore IS6. The profound inactivation rate slowing caused by $\text{Ca}_v\beta_{2a}$ requires anchoring of the N terminus to the membrane by palmitoylation²⁸. Orientation of the AID helix towards IS6 places the N-terminal membrane anchor on the periphery of the α_1 subunit and suggests that $\text{Ca}_v\beta_{2a}$ slows channel inactivation by restricting the movement of the IS6 transmembrane domain. $\text{Ca}_v\beta$ s have a deep groove between the SH3 and

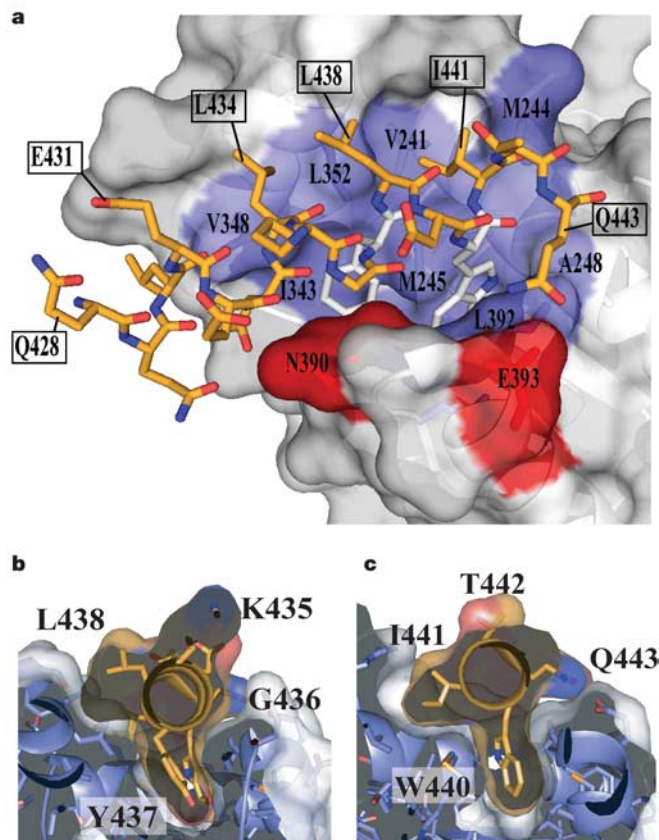


Figure 4 AID–ABP interactions. **a**, Surface representation of the $\text{Ca}_v\beta_{2a}$ ABP, bound to the AID. The AID (gold) is shown in stick representation. Y437 and W440 are white. $\text{Ca}_v\beta_{2a}$ residues that contribute hydrophobic (blue) and hydrogen bond (red) side-chain contacts to the AID are labelled. Select residues of the AID are labelled to orient the reader. **b**, **c**, Slices through the AID–ABP interaction at AID positions Y437 and W440 (gold). Labels indicate the AID residues.

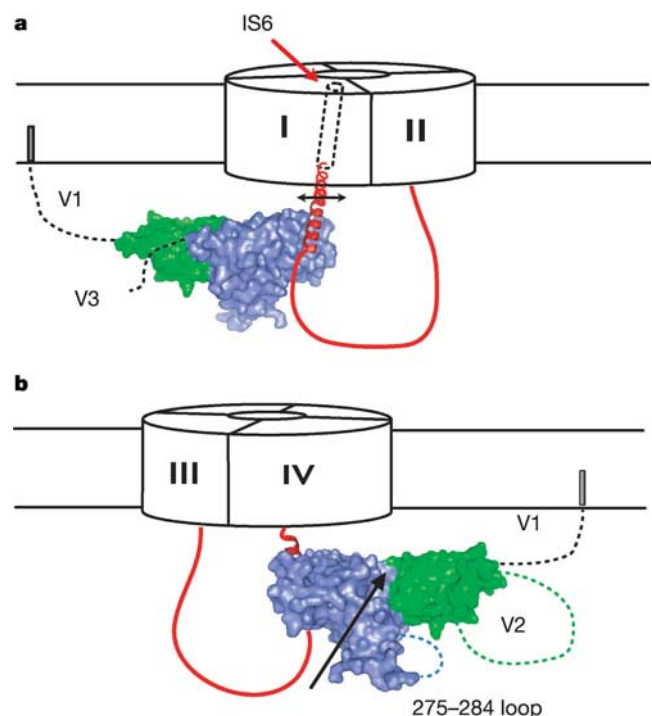


Figure 5 Cartoon of proposed model for how $\text{Ca}_v\beta$ affects $\text{Ca}_v\alpha_1$ gating. **a**, Orientation of $\text{Ca}_v\beta_{2a}$ with respect to the I–II loop (red), pore-forming subunit, and connection to IS6. In $\text{Ca}_v\beta_{2a}$, variable region 1 (V1) is tethered to the membrane. The I–II loop between the AID N terminus and IS6 is depicted as a helix. $\text{Ca}_v\beta_{2a}$ SH3 and nucleotide kinase domains are coloured green and blue, respectively. The arrow indicates that $\text{Ca}_v\beta$ couples to IS6 movements (rotations, translations or both). **b**, View from the opposite side of **a**. The groove between SH3 and nucleotide kinase domains (demarcated by the arrow) and two flexible $\text{Ca}_v\beta_{2a}$ regions, the 275–284 loop and variable region 2 (V2), are on the same $\text{Ca}_v\beta$ face, opposite the ABP. These regions may interact with other pore-forming subunit cytoplasmic domains.

nucleotide kinase domains that is on the same face as the V2 domain. These features may be used to engage other cytoplasmic parts of the channel^{29,30} and allow $\text{Ca}_v\beta$ to couple motions in other channel domains directly to IS6 through AID attachment.

The structures presented here represent the first high-resolution view of any part of the voltage-gated calcium channel, and provide an important step towards understanding the detailed molecular mechanism models for how $\text{Ca}_v\beta$ s function. This work strongly suggests that $\text{Ca}_v\beta$ affects Ca_v gating properties by directly influencing conformational changes that are likely to occur in the channel pore¹⁰. □

Methods

The crystal structures of recombinant $\text{Ca}_v\beta_{2a}$ and $\text{Ca}_v\beta_{2a}$ -AID complexes were solved to resolutions of 1.97 Å and 2.00 Å, respectively. The final R/R_{free} values are 18.55%/21.32% for $\text{Ca}_v\beta_{2a}$ and 19.97%/24.15% for the $\text{Ca}_v\beta_{2a}$ -AID complex. Figures were prepared with PyMOL, MOLSCRIPT and RASTER3D. The experimental details for protein expression, purification, crystallization, structure solution, statistics of data collection, phasing and refinement are available as Supplementary Information.

Received 1 March; accepted 23 April 2004; doi:10.1038/nature02588.

Published online 12 May 2004.

1. Catterall, W. A. Structure and regulation of voltage-gated Ca^{2+} channels. *Annu. Rev. Cell Dev. Biol.* **16**, 521–555 (2000).
2. Dolphin, A. C. β -subunits of voltage-gated calcium channels. *J. Bioenerg. Biomembr.* **35**, 599–620 (2003).
3. Pragnell, M. *et al.* Calcium channel β -subunit binds to a conserved motif in the I–II cytoplasmic linker of the α_1 -subunit. *Nature* **368**, 67–70 (1994).
4. Dolphin, A. C. G protein modulation of voltage-gated calcium channels. *Pharmacol. Rev.* **55**, 607–627 (2003).
5. Hering, S. β -Subunits: fine tuning of Ca^{2+} channel block. *Trends Pharmacol. Sci.* **23**, 509–513 (2002).
6. Bichet, D. *et al.* The I–II loop of the Ca^{2+} channel α_1 subunit contains an endoplasmic reticulum retention signal antagonized by the β subunit. *Neuron* **25**, 177–190 (2000).
7. Beguin, P. *et al.* Regulation of Ca^{2+} channel expression at the cell surface by the small G-protein Kir/Gem . *Nature* **411**, 701–706 (2001).
8. De Waard, M., Scott, V. E., Pragnell, M. & Campbell, K. P. Identification of critical amino acids involved in $\alpha_1\beta$ interaction in voltage-dependent Ca^{2+} channels. *FEBS Lett.* **380**, 272–276 (1996).
9. Stotz, S. C., Hamid, J., Spaetgens, R. L., Jarvis, S. E. & Zamponi, G. W. Fast inactivation of voltage-dependent calcium channels. A hinged-lid mechanism? *J. Biol. Chem.* **275**, 24575–24582 (2000).
10. Zhang, J. F., Ellinor, P. T., Aldrich, R. W. & Tsien, R. W. Molecular determinants of voltage-dependent inactivation in calcium channels. *Nature* **372**, 97–100 (1994).
11. Dimitratos, S. D., Woods, D. F., Stathakis, D. G. & Bryant, P. J. Signaling pathways are focused at specialized regions of the plasma membrane by scaffolding proteins of the MAGUK family. *Bioessays* **21**, 912–921 (1999).
12. Tavares, G. A., Panepucci, E. H. & Brunger, A. T. Structural characterization of the intramolecular interaction between the SH3 and guanylate kinase domains of PSD-95. *Mol. Cell* **8**, 1313–1325 (2001).
13. McGee, A. W. *et al.* Structure of the SH3–guanylate kinase module from PSD-95 suggests a mechanism for regulated assembly of MAGUK scaffolding proteins. *Mol. Cell* **8**, 1291–1301 (2001).
14. Paarmann, L., Spangenberg, O., Lavie, A. & Konrad, M. Formation of complexes between Ca^{2+} -calmodulin and the synapse-associated protein SAP97 requires the SH3 domain–guanylate kinase domain-connecting HOOK region. *J. Biol. Chem.* **277**, 40832–40838 (2002).
15. Witcher, D. R., De Waard, M., Liu, H., Pragnell, M. & Campbell, K. P. Association of native Ca^{2+} channel β subunits with the α_1 subunit interaction domain. *J. Biol. Chem.* **270**, 18088–18093 (1995).
16. De Waard, M., Pragnell, M., & Campbell, K. P. Ca^{2+} channel regulation by a conserved β subunit domain. *Neuron* **13**, 495–503 (1994).
17. Berrou, L., Klein, H., Bernatchez, G. & Parent, L. A specific tryptophan in the I–II linker is a key determinant of β -subunit binding and modulation in $\text{Ca(V)}2.3$ calcium channels. *Biophys. J.* **83**, 1429–1442 (2002).
18. Opatowsky, Y., Chomsky-Hecht, O., Kang, M. G., Campbell, K. P. & Hirsch, J. A. The voltage-dependent calcium channel β subunit contains two stable interacting domains. *J. Biol. Chem.* **278**, 52323–52332 (2003).
19. Kochegarov, A. A. Pharmacological modulators of voltage-gated calcium channels and their therapeutic application. *Cell Calcium* **33**, 145–162 (2003).
20. De Waard, M. *et al.* Direct binding of G-protein $\beta\gamma$ complex to voltage-dependent calcium channels. *Nature* **385**, 446–450 (1997).
21. Zhang, J. F., Ellinor, P. T., Aldrich, R. W. & Tsien, R. W. Multiple structural elements in voltage-dependent Ca^{2+} channels support their inhibition by G proteins. *Neuron* **17**, 991–1003 (1996).
22. Qin, N., Platano, D., Olcese, R., Stefani, E. & Birnbaumer, L. Direct interaction of $\text{G}\beta\gamma$ with a C-terminal $\text{G}\beta\gamma$ -binding domain of the Ca^{2+} channel α_1 subunit is responsible for channel inhibition by G protein-coupled receptors. *Proc. Natl Acad. Sci. USA* **94**, 8866–8871 (1997).
23. Ivanina, T., Blumenstein, Y., Shistik, E., Barzilai, R. & Dascal, N. Modulation of L-type Ca^{2+} channels by $\text{G}\beta\gamma$ and calmodulin via interactions with N and C termini of $\alpha_1\text{C}$. *J. Biol. Chem.* **275**, 39846–39854 (2000).
24. Miyazawa, A., Fujiyoshi, Y. & Unwin, N. Structure and gating mechanism of the acetylcholine receptor pore. *Nature* **424**, 949–955 (2003).
25. Jiang, Y. *et al.* The open pore conformation of potassium channels. *Nature* **417**, 523–526 (2002).
26. Herlitze, S., Hockerman, G. H., Scheuer, T. & Catterall, W. A. Molecular determinants of inactivation and G protein modulation in the intracellular loop connecting domains I and II of the calcium channel $\alpha_1\text{A}$ subunit. *Proc. Natl Acad. Sci. USA* **94**, 1512–1516 (1997).
27. Berrou, L., Bernatchez, G. & Parent, L. Molecular determinants of inactivation within the I–II linker of $\alpha_1\text{E}$ ($\text{CaV}2.3$) calcium channels. *Biophys. J.* **80**, 215–228 (2001).

28. Restituito, S. *et al.* The $\beta_2\text{a}$ subunit is a molecular groom for the Ca^{2+} channel inactivation gate. *J. Neurosci.* **20**, 9046–9052 (2000).
29. Walker, D. *et al.* A new β subtype-specific interaction in $\alpha_1\text{A}$ subunit controls P/Q-type Ca^{2+} channel activation. *J. Biol. Chem.* **274**, 12383–12390 (1999).
30. Walker, D., Bichet, D., Campbell, K. P. & De Waard, M. A β_4 isoform-specific interaction site in the carboxyl-terminal region of the voltage-dependent Ca^{2+} channel $\alpha_1\text{A}$ subunit. *J. Biol. Chem.* **273**, 2361–2367 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. M. Berger, K. Brejc, D. Fass, D. Julius, E. A. Lumpkin and B. A. Schulman for comments on the manuscript; J. Holton at beamline 8.3.1 at the Advanced Light Source for help with data collection; R. W. Tsien and D. T. Yue for the calcium channel clones; and members of the Minor laboratory for support at all stages of this work. This work was supported by awards to D.L.M. from the McKnight Foundation for Neuroscience, the March of Dimes Basil O'Connor Scholar program, the Alfred P. Sloan Foundation, and the Rita Allen Foundation. D.L.M. is a McKnight Foundation Scholar, an Alfred P. Sloan Research Fellow and a Rita Allen Foundation Scholar.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.L.M. (minor@itsa.ucsf.edu). Coordinates and structure factors have been deposited in the Protein Data Bank under accession codes 1T0H and 1T0J.

Structural basis of the $\alpha_1\beta$ subunit interaction of voltage-gated Ca^{2+} channels

Yu-hang Chen, Ming-hui Li, Yun Zhang, Lin-ling He, Yoichi Yamada, Aileen Fitzmaurice, Yang Shen, Hailong Zhang, Liang Tong & Jian Yang

Department of Biological Sciences, Columbia University, New York, New York 10027, USA

High-voltage-activated Ca^{2+} channels are essential for diverse biological processes. They are composed of four or five subunits, including α_1 , $\alpha_2\text{-}\delta$, β and γ (ref. 1). Their expression and function are critically dependent on the β -subunit, which transports α_1 to the surface membrane and regulates diverse channel properties^{2–4}. It is believed^{3–6} that the β -subunit interacts with α_1 primarily through the β -interaction domain (BID), which binds directly to the α -interaction domain (AID) of α_1 ⁷; however, the molecular mechanism of the $\alpha_1\beta$ interaction is largely unclear. Here we report the crystal structures of the conserved core region of β_3 , alone and in complex with AID, and of β_4 alone. The structures show that the β -subunit core contains two interacting domains: a Src homology 3 (SH3) domain and a guanylate kinase (GK) domain. The AID binds to a hydrophobic groove in the GK domain through extensive interactions, conferring extremely high affinity between α_1 and β -subunits^{4,8}. The BID is essential both for the structural integrity of and for bridging the SH3 and GK domains, but it does not participate directly in binding α_1 . The presence of multiple protein-interacting modules in the β -subunit opens a new dimension to its function as a multi-functional protein.

Four different types of β -subunits ($\beta_1\text{--}\beta_4$) have been identified, each with multiple splicing variants (for review, see refs 2–4). Previous studies show that the β -subunit contains five regions, with the second and forth being highly conserved (68–92% identity) and the others highly variable among the four β -subfamilies^{5,9,10}. The three middle regions form a core that is able to confer most of the functional properties of the full-length β -subunit^{5,9}. Molecular modelling indicates that the second and forth regions are homologous to canonical SH3 domains and the yeast GK domain, respectively, and based on this homology, a structural model of

these two regions was developed¹⁰. This homology model opened the possibility that the β -subunit contains several common protein-interaction modules; however, the validity of this model is unclear and the arrangement of the presumed SH3 and GK domains is unknown. Furthermore, the presumed interaction between BID and AID remains unconfirmed.

We first crystallized the core region encompassing amino acids 16–366 of rat β_3 (Fig. 1a, referred to as the β_3 core hereafter) and determined its structure at 2.3 Å resolution (Supplementary Fig. 1a) by single-wavelength anomalous diffraction with selenomethionine-substituted protein (Supplementary Table 1)¹¹. The current structural model contains residues 31–140, 167–225 and 245–361, and 397 water molecules. Other residues are not included owing to disorder. We also crystallized the β_3 core in complex with a 49-amino-acid (residues A415–Q463) segment of the intracellular

loop connecting the first two homologous repeats of $\text{Ca}_v1.2c$, an α_1 subunit encoding rat brain L-type Ca^{2+} channels (Fig. 1c). This segment contains the 18-amino-acid AID that has been shown to bind the β -subunit directly⁷. The structure of the complex (Fig. 2) was determined at 2.6 Å resolution by the molecular replacement method. The current model includes residues 38–136 and 167–361 of β_3 and 422–446 of $\text{Ca}_v1.2c$. In addition, we crystallized the same region of rat β_4 and determined its structure at 3.0 Å resolution by the molecular replacement method. The current model includes residues 78–168, 208–266 and 286–402.

The β_3 core is comprised of three regions: an SH3 domain (residues 60–120 and 170–175), a HOOK region (residues 121–169, following the nomenclature of a homologous region in PSD95 (ref. 12)) and a GK domain (residues 176–360) (Fig. 2). The structure of the SH3 and GK domains is significantly different

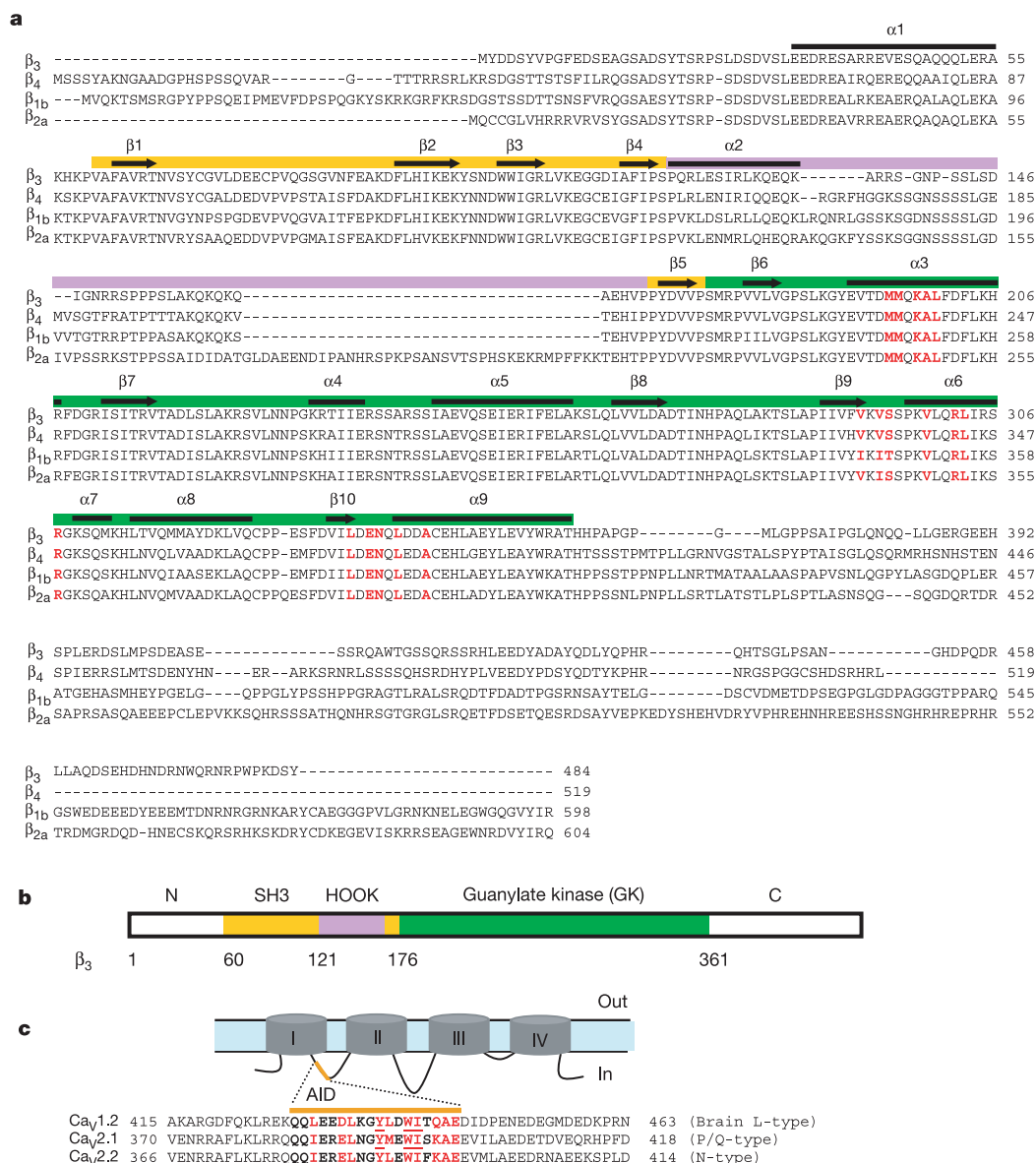


Figure 1 Sequence alignment and domain organization of Ca^{2+} channel α_1 and β -subunits. **a**, Amino acid sequence alignment of β_3 (GenBank accession number M88751), β_4 (L02315), β_{1b} (NP-000714) and β_{2a} (M80545). Gold indicates the SH3 domain, purple the HOOK region and green the GK domain. Secondary structure elements in the β_3 subunit are indicated in the top line (arrows for β -sheets and solid lines for α -helices). Residues involved in interactions with AID are marked in red. **b**, Schematic

diagram of the domain organization of β_3 based on the crystal structure. **c**, Schematic domain organization of α_1 and alignment of the 49-amino-acid segment from $\text{Ca}_v1.2c$ (M67515), $\text{Ca}_v2.1$ (X57477) and $\text{Ca}_v2.2$ (D14157). Residues forming the AID are in bold and those involved in interactions with the β -subunit are in red, with the three most critical residues underlined.

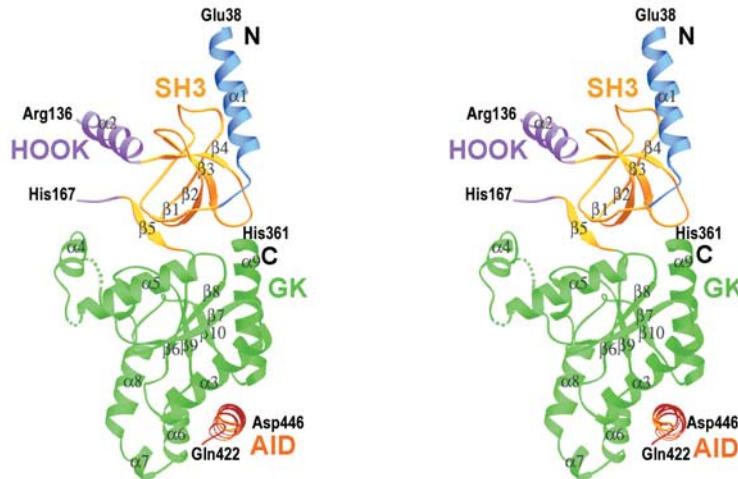


Figure 2 Crystal structure of the β_3 core in complex with its α_1 binding partner (AID). The β_3 core can be divided into three regions: an SH3 domain (residues 60–120 and 170–175, gold), a HOOK region (residues 121–169, purple) and a GK domain (residues 176–360, green). Residues 226–244 (forming the α_4 helix) are disordered in this

molecule but are well resolved in another one in the same asymmetric unit. Residues 422–446 of Ca ν 1.2c containing the entire AID are coloured in orange. This and all subsequent ribbon diagrams were generated with the program RIBBONS²⁸.

from that predicted by homology modelling¹⁰, and the two domains interact intramolecularly. The AID (residues Q428–E445) forms an α -helix and binds to a hydrophobic groove in the GK domain located on the opposite side of the SH3 domain (Figs 2 and 3a, c). This groove is walled by helices α_3 , α_6 and α_9 , and the loop connecting β_9 and α_6 and β_{10} and α_9 (Figs 2 and 3c). The interactions between AID and the β -subunit are predominantly hydrophobic, but also include hydrogen bonds and an ion pair (Fig. 3b, c). These extensive interactions are probably the basis for the high-affinity (~ 20 nM) binding between α_1 and β (refs 5, 8). Most of the residues involved in the interactions are conserved in both α_1 and β (Fig. 1a, c). Previous mutagenesis studies indicate that three residues in AID (Y437, W440 and I441 in Ca ν 1.2c; Fig. 1c) are particularly important for α_1 – β interaction^{6,7}. The structure shows that these residues exhibit the most extensive interactions with the β -subunit (Fig. 3b, c), agreeing well with their functional importance.

Notably, the structure reveals that the BID (residues K163–T193 in β_3) is not directly involved in binding AID (Fig. 2). Instead, it

has an essential structural role. It spans all three regions of the SH3–HOOK–GK motif and contains two β -strands (β_5 and β_6), which constitute an integral part of the SH3 and GK domains (Supplementary Fig. 2). Thus, the BID is crucial for the intramolecular association of the SH3 and GK domains as well as for their structural integrity. Mutations on the BID that were previously found to abolish α_1 – β interactions^{5,6} probably altered the folding and/or structure of the β -subunit, and current enhancement by a 35-amino-acid peptide containing the BID⁵ is probably a non-specific effect (Supplementary Fig. 3).

The structure of the *apo*- β_3 core is very similar to that of the complex between the β_3 core and AID (Fig. 4a). Superposition of the SH3 and GK domains in the two structures separately shows a root-mean-square deviation (r.m.s.d.) of 0.44 Å (88 C α atoms) and 0.51 Å (165 C α atoms), respectively. The small difference in the quaternary arrangement of the SH3 and GK domains (Fig. 4a) is probably due to different crystal packing. In and near the AID-binding site, however, there is a small but discernible backbone movement, with helices α_6 , α_7 and α_9 moving away from the

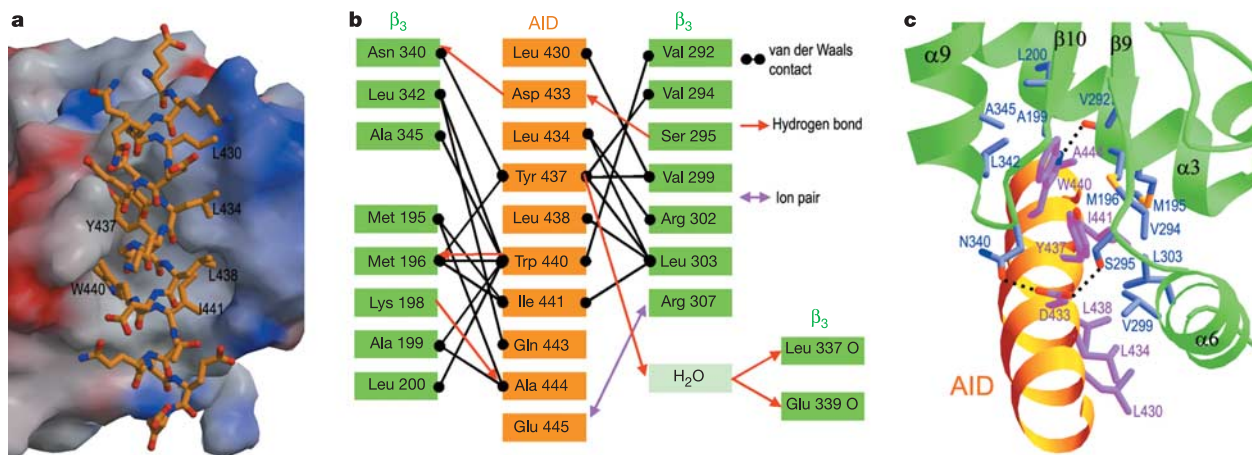


Figure 3 Interactions between β_3 and AID. **a**, GRASP²⁹ surface representation of the β_3 core–AID interface. Red and blue represent negative and positive electrostatic potential, respectively. Selected hydrophobic residues are labelled in AID (stick representation).

b, Specific interactions between β_3 and AID. **c**, Close-up of the interface between β_3 and AID. Most, but not all, residues involved in the interactions are shown. The colour scheme in **c** follows that of Fig. 2.

binding pocket in the complex (Fig. 4b). The side chains of six residues show even larger movements. In particular, the side chains of M195, M196 and L342 extend into the AID-binding pocket in the *apo*-structure but swing away in the complex, as if to make room for AID (Fig. 4b). Overall, these movements are local and binding of AID does not cause a significant and far-reaching structural change of the β -subunit core.

The structure of the β_4 core is similar to that of the β_3 core (Fig. 4c). The r.m.s.d. is 0.97 Å when 272 C α atoms are superimposed in the two structures. This is not surprising considering that the SH3 and GK domains of these two subunits share a 75% and 89% identity.

To understand the functions of the β -subunit SH3 and GK domains, it is of interest to compare them to other known structures. Canonical SH3 domains mediate specific protein–protein interactions by binding to PXXP-containing motifs in target proteins¹³. They are composed of five sequential β -strands (β 1– β 5) arranged into two orthogonally packed sheets^{13,14}. The β -subunit SH3 domain (β -SH3) shares this overall structure. Comparison of β_3 -SH3 and a canonical SH3 domain (of Fyn tyrosine kinase) shows a high degree of overlap in the β -strands (Fig. 5a); however, the connecting loops diverge significantly. In particular, the loop between β -strands β 1 and β 2 (that is, RT-Src loop), where several PXXP-motif-interacting residues reside, is considerably longer in β -SH3. Furthermore, in β -SH3 a long and flexible HOOK region

separates β -strands β 4 and β 5, replacing a conserved, short 3_{10} helix in canonical SH3 domains. Thus, β -SH3 has a split configuration, with a separate region being inserted between the first four β -strands (β 1– β 4) and the fifth one (β 5).

Despite these structural differences, several amino acids critical for binding PXXP motifs are conserved in β -SH3 (Fig. 5a). These include P119, W103, Y70, V68, and F117 in β_3 , and they are among the most conserved residues in canonical SH3 domains¹⁴. They form an aromatic patch on the opposite side of the SH3–GK interface. This binding site, however, is shielded by the long RT-Src loop and the α 2 helix (Figs 2 and 5a). Simulated docking of a PXXP peptide to this site reveals multiple clashes between the peptide and these two regions (data not shown). Thus, β -SH3 is unlikely to interact with PXXP motifs unless the RT-Src loop and α 2 helix move out of the way. Nevertheless, the conservation of this ligand-binding site raises the possibility that β -SH3 may function as a protein-interaction module. In this context, it is interesting to note that multiple PXXP motifs are present in the cytoplasmic domains of the α_1 subunit of high-voltage-activated Ca²⁺ channels.

The HOOK region has a variable length and a relatively low overall amino acid identity among the β -families (Fig. 1a); however, the 13 amino-terminal residues share very high similarity and form an α -helix (α 2) (Fig. 2). The remaining residues are variable and are disordered in both β_3 core and β_4 core structures.

The overall structure of the β -subunit GK domain (β -GK) is similar to that¹⁵ of the yeast GK domain (Fig. 5b); however, many regions diverge markedly. Of the 186 C α atoms, 85 are >2.5 Å apart in the two structures. Notably, β -GK is clearly catalytically inactive because many residues critical for GMP- and ATP-binding¹⁵ do not exist in β -GK.

Interestingly, the SH3–HOOK–GK motif is also a common structural signature of a family of membrane-associated GK homologues (MAGUKs), which also contain one or more PDZ domains. Many members of this family, including PSD95, SAP97 and CASK are concentrated at synapses and have a central role in clustering and organizing both pre- and post-synaptic ion channels and neurotransmitter receptors¹⁶. The presence of a common SH3–HOOK–GK motif in β -subunits and PSD95 family proteins, and their co-localization at synapses, raise an intriguing question of whether they cross-interact with each other's partners. Besides the α_1 subunit, the only other known group of proteins that interacts directly with the β -subunit is GEM/Rem/Rad, a family of Ras-related small G proteins that inhibit Ca²⁺ channel expression at the cell surface when expressed in several cell types^{17,18}. Conversely, although PSD95 interacts with a number of voltage- and ligand-gated channels¹⁶, there is as yet no report of its interaction with high-voltage-activated Ca²⁺ channels. To examine whether the PSD95 SH3–HOOK–GK motif interacts with high-voltage-activated Ca²⁺ channels, we expressed in *Xenopus* oocytes Cav2.1 (the α_1 subunit encoding P/Q-type channels), $\alpha_2\delta$ and either wild-type β_3 , β_3 core, β_3 -GK, PSD95 SH3–HOOK–GK or PSD95-GK. Surface expression of exogenous Ca²⁺ channels was monitored by measuring the whole-oocyte current. Co-expression of β_3 core or β_3 -GK increased the current to a similar extent as did wild-type β_3 (Fig. 5c). On the other hand, PSD95 SH3–HOOK–GK and PSD95-GK failed to increase the current (Fig. 5c), suggesting that they do not interact with Cav2.1. Comparison of the structure of PSD95-GK and β -GK clearly reveals why: the AID-binding pocket is completely occluded in PSD95-GK (Fig. 5d).

The split configuration of β -SH3 is similar to that of PSD95-SH3^{12,19}. These two SH3 domains show a high degree of overlap in the β -strands but significant deviations in the connecting loops (Fig. 5e). A marked difference is that in PSD95-SH3 the β -strand β 5 is significantly longer and forms an anti-parallel β -sheet with an additional β -strand (β F) formed by the carboxy terminus of the GK domain, an interaction critical for the intramolecular assembly of the PSD95 SH3–HOOK–GK module^{12,19}. This extra β -sheet does

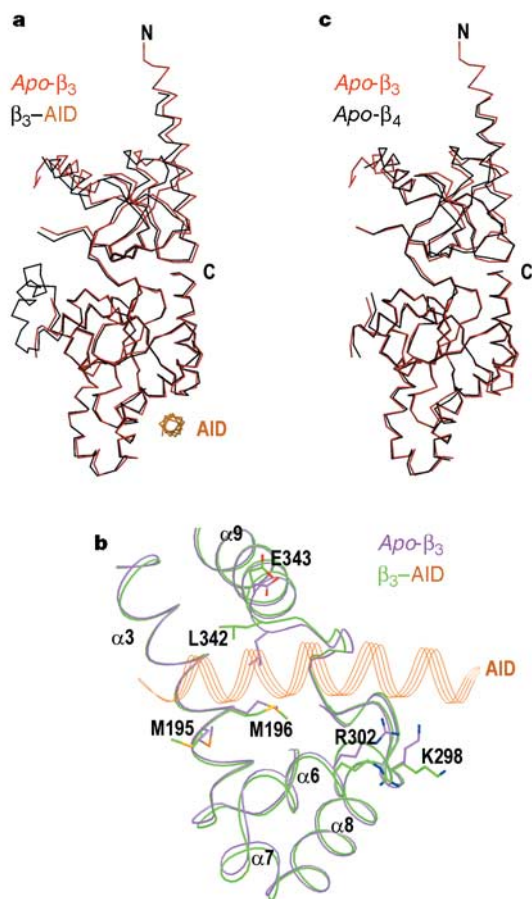


Figure 4 Structural comparisons of the β -subunit core. **a**, Superposition of the *apo*- β_3 core (red) and β_3 core–AID complex (black). This and all subsequent trace diagrams were produced with the program Molscript²⁰. **b**, Close-up of the AID-binding site of *apo*- β_3 core (magenta) and β_3 core–AID complex (green). **c**, Superposition of β_3 core (red) and β_4 core (black). In **a** and **c**, only the GK domains were superimposed. Stereo views of **a** and **c** are in Supplementary Fig. 1.

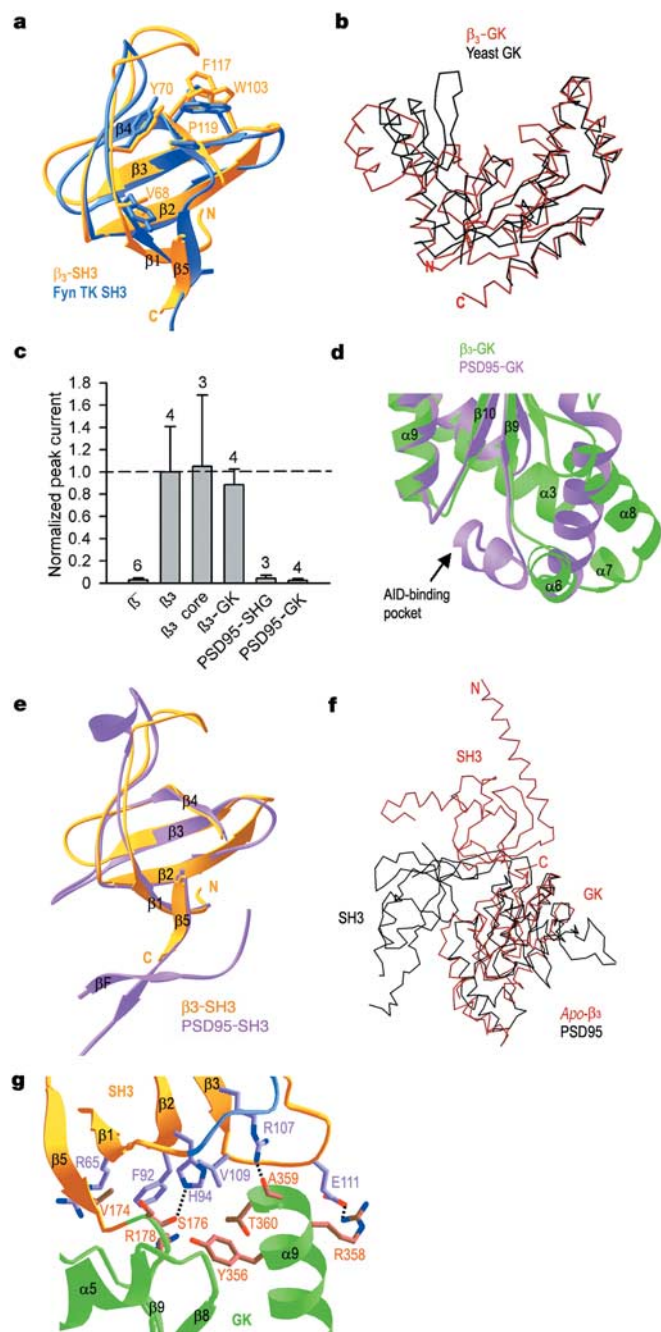


Figure 5 Structural and functional comparisons of the β -subunit with other proteins. **a**, Superposition of the SH3 domain of *apo*- β_3 (gold) and Fyn tyrosine kinase (Protein Data Bank code 1sfh, blue). Side chains conserved for ligand binding are shown but are labelled only for β_3 . **b**, Superposition of the GK domain of β_3 (red) and yeast GK (1EX6, black). **c**, Whole-oocyte current measured in oocytes injected with $\text{Ca}_v2.1$ and $\alpha_2\delta$ only (β^-) or together with the indicated construct. β_3 -GK represents amino acids M177–G366 of β_3 ; PSD95-SHG represents the SH3–HOOK–GK motif (amino acids G473–L767) of PSD95; and PSD95-GK represents the GK domain (amino acids H575–S755) of PSD95. Current was normalized by the mean current obtained with wild-type β_3 . Data are presented as mean $\pm$ s.d. with the number of observations indicated above the error bars. **d**, Superposition of the GK domain of *apo*- β_3 (green) and PSD95 (1KJW, pink). Only the region around the AID-binding site is shown. Secondary structures are labelled only for β_3 . **e**, Superposition of the SH3 domain of *apo*- β_3 (gold) and PSD95 (1KJW, pink). Note that the β -sheet formed by β_5 and β_6 in PSD95 does not exist in β_3 . **f**, Comparison of the SH3–HOOK–GK motif of *apo*- β_3 (red) and PSD95 (1KJW, black). The GK domains were superimposed, leaving the HOOK region and SH3 domain unconstrained. **g**, Close-up of the SH3 and GK interface in *apo*- β_3 core.

not exist in β -subunits. Another difference is that in PSD95-SH3, the PXXP-motif-binding site is not obscured by the RT-Src loop and hence is accessible (Fig. 5e). This is consistent with the observation that the SH3 domain of MAGUKs can associate directly with PXXP-motif-containing proteins^{20,21}.

The relative arrangement of and interactions between the SH3 and GK domains are also different in PSD95 and β -subunits (Fig. 5f). These two domains are brought into close proximity by interactions between β -strands β_1 and β_5 in both proteins. They associate with each other through a limited number of hydrogen bonds and van der Waals contacts in β -subunits (Fig. 5g), but in PSD95 they are joined together by much more extensive interactions, including the β -sheet formed between β -strands β_5 and β_6 ^{12,19} (Fig. 5e). As a result, the SH3 and GK domains 'crouch' together in a U-shaped conformation in PSD95 whereas they relax into a more extended conformation in β -subunits (Fig. 5f). The functional importance of this intramolecular interaction in the β -subunit remains to be examined.

Our studies show that the Ca^{2+} channel β -subunit contains two common protein-interaction domains (SH3 and GK) and interacts with the α_1 subunit not through the BID but via a hydrophobic groove in the GK domain. This interaction anchors β to α_1 . It is interesting to note that AID occupies only a very small area of the SH3–HOOK–GK motif and much of the surface of this motif is free (Fig. 2 and Supplementary Fig. 1d) and available for interactions with other regions of α_1 or other proteins. In MAGUK proteins, such as PSD95, each region in the SH3–HOOK–GK motif has been found to interact with a number of proteins^{12,16,19}. It is tempting to speculate that these protein-interacting modules in the Ca^{2+} channel β -subunit may have roles in addition to chaperoning α_1 and modulating its properties. □

Methods

Molecular biology, protein expression and purification

All DNA constructs were made by polymerase chain reaction (PCR) and were confirmed by sequencing.

The gene encoding amino acids G16–G366 of rat brain β_3 subunit (β_3 core) was subcloned into pET-26b and expressed in BL21(DE3) at 25 °C. After cell lysis, the soluble protein was purified by cobalt-resin affinity and gel-filtration chromatography. The protein was concentrated to 10 mg ml⁻¹ in a buffer containing 10 mM Na-HEPES (pH 7.2), 250 mM NaCl, 2 mM dithiothreitol, and 1% (v/v) glycerol and was stored at -80 °C.

For production of selenomethionyl proteins, the expression construct was transformed into the methionine auxotroph B834(DE3). Bacteria were grown in defined LeMaster media and the protein was purified using the same protocol as for the native protein. The C-terminal His tag was not removed for crystallization.

The gene encoding residues G49–T407 of rat brain β_4 subunit (β_4 core) was subcloned into pET-28a. The selenomethionyl protein was first purified by cobalt-resin affinity column, and the N-terminal His tag was removed by incubation with thrombin (1 U mg⁻¹ protein) for 12 h at 25 °C. Thrombin-cleaved protein was further purified by gel-filtration chromatography, concentrated to 10 mg ml⁻¹ in the same buffer as mentioned above, and stored at -80 °C.

A 49-amino-acid segment of the rat brain $\text{Ca}_v1.2c$ subunit (residues Ala415–Gln 463) containing the AID domain was subcloned into pACYCduet-1 and co-expressed with β_3 core (in pET-26b) in BL21(DE3) at 25 °C. The β_3 core–AID complex was purified, concentrated and stored as described above, with the exception that the buffer contained 10 mM Tris-HCl (pH 8.0) instead of 10 mM Na-HEPES (pH 7.2).

Protein crystallization

Crystallization of the selenomethionyl β_3 core protein was carried out using the hanging-drop vapour diffusion method at 21 °C. The reservoir solution contained 20 mM Na-HEPES (pH 7.2), 250 mM NaCl and 15–25% (w/v) polyethylene glycol 3350, and microseeding was used to produce larger crystals. The crystals were transferred to a cryoprotectant solution containing 20 mM Na-HEPES (pH 7.2), 20% (w/v) polyethylene glycol 3350 and 10% (v/v) ethylene-glycol, and flash-frozen in liquid nitrogen for data collection at 100 K. Crystals of selenomethionyl β_4 core protein were obtained by the same method and were grown against a reservoir solution containing 20 mM Na-HEPES (pH 7.2), 250 mM NaCl and 5–10% (w/v) polyethylene glycol 3350. Crystals of β_3 core–AID were obtained by vapour diffusion against a reservoir solution containing 20 mM Tris-HCl (pH 8.0), 150 mM MgCl₂ and 15% (w/v) polyethylene glycol 3350, and microseeding was used to produce larger crystals.

Data collection and processing

A single-wavelength anomalous diffraction data set for the β_3 core crystal was collected to

2.3 Å resolution. The crystal belongs to space group $P2_1$, with cell dimensions of $a = 57.3$ Å, $b = 66.4$ Å, $c = 90.0$ Å and $\beta = 103.1^\circ$. For the β_4 core crystal, a 3.0 Å resolution data set was collected. The crystal also belongs to space group $P2_1$, but is non-isomorphous with that of β_3 core, with cell dimensions of $a = 60.3$ Å, $b = 83.2$ Å, $c = 72.3$ Å and $\beta = 94.9^\circ$. The data set for the β_3 core-AID crystal was collected to 2.6 Å resolution. It belongs to space group $C2$, with cell dimensions of $a = 252.3$ Å, $b = 69.0$ Å, $c = 60.7$ Å and $\beta = 96.7^\circ$. There are two molecules in the asymmetric unit for all three crystals, giving a V_m (Matthew's coefficient) of $2.0 \text{ Å}^3 \text{ Da}^{-1}$, $2.3 \text{ Å}^3 \text{ Da}^{-1}$ and $2.9 \text{ Å}^3 \text{ Da}^{-1}$, respectively. All diffraction data were collected at the X4A beamline of the Brookhaven National Laboratory. The diffraction images were processed and scaled with the HKL package²². The data processing statistics are summarized in Supplementary Table 1.

Structure determination and refinement

The locations of Se atoms in the β_3 core crystal were determined with the program SOLVE²³. Reflection phases to 2.3 Å resolution were calculated and improved with the same program, which also automatically located 50% of the residues in the molecule. The full atomic model was built into the electron density with the program O²⁴. Structure refinement was carried out with the program CNS²⁵. The structures of the β_3 core-AID complex and β_4 core were determined by the molecular replacement method with the program COMO²⁶ using the β_3 core structure as the model, and were refined as described above. The statistics on structural refinement are summarized in Supplementary Table 1. The current models fit the electron density map well and all main-chain dihedral angles are located in favourable regions on the Ramachandran plot (Supplementary Fig. 2).

Oocyte expression and electrophysiology

All constructs for oocyte expression were subcloned into variants of pGEMHE. cRNAs were synthesized *in vitro* and injected in various combinations into *Xenopus* oocytes, which were obtained and maintained as described²⁷. Two-electrode voltage-clamp recordings were performed as described²⁷. Data are presented as mean $\pm$ s.d. (number of observations).

Received 5 March; accepted 10 May 2004; doi:10.1038/nature02641.

Published online 30 May 2004.

1. Catterall, W. A. Structure and regulation of voltage-gated Ca^{2+} channels. *Annu. Rev. Cell Dev. Biol.* **16**, 521–555 (2000).
2. Birnbaumer, L. *et al.* Structures and functions of calcium channel beta subunits. *J. Bioenerg. Biomembr.* **30**, 357–375 (1998).
3. Arikath, J. & Campbell, K. P. Auxiliary subunits: essential components of the voltage-gated calcium channel complex. *Curr. Opin. Neurobiol.* **13**, 298–307 (2003).
4. Dolphin, A. C. β subunits of voltage-gated calcium channels. *J. Bioenerg. Biomembr.* **55**, 607–627 (2003).
5. De Waard, M., Pragnell, M. & Campbell, K. P. Ca^{2+} channel regulation by a conserved beta subunit domain. *Neuron* **13**, 495–503 (1994).
6. De Waard, M., Scott, V. E., Pragnell, M. & Campbell, K. P. Identification of critical amino acids involved in $\alpha 1$ - β interaction in voltage-dependent Ca^{2+} channels. *FEBS Lett.* **380**, 272–276 (1996).
7. Pragnell, M. *et al.* Calcium channel β -subunit binds to a conserved motif in the I-II cytoplasmic linker of the $\alpha 1$ -subunit. *Nature* **368**, 67–70 (1994).
8. Canti, C. *et al.* Evidence for two concentration-dependent processes for β -subunit effects on $\alpha 1\text{B}$ calcium channels. *Biophys. J.* **81**, 1439–1451 (2001).
9. Opatowsky, Y., Chomsky-Hecht, O., Kang, M. G., Campbell, K. P. & Hirsch, J. A. The voltage-dependent calcium channel beta subunit contains two stable interacting domains. *J. Biol. Chem.* **278**, 52323–52332 (2003).
10. Hanlon, M. R., Berrow, N. S., Dolphin, A. C. & Wallace, B. A. Modelling of a voltage-dependent Ca^{2+} channel beta subunit as a basis for understanding its functional properties. *FEBS Lett.* **445**, 366–370 (1999).
11. Hendrickson, W. A. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**, 51–58 (1991).
12. Tavares, G. A., Panepucci, E. H. & Brunger, A. T. Structural characterization of the intramolecular interaction between the SH3 and guanylate kinase domains of PSD-95. *Mol. Cell* **8**, 1313–1325 (2001).
13. Zarrinpar, A., Bhattacharyya, R. P. & Lim, W. A. The structure and function of proline recognition domains. *Sci. STKE* **2003**, RE8 (2003).
14. Larson, S. M. & Davidson, A. R. The identification of conserved interactions within the SH3 domain by alignment of sequences and structures. *Protein Sci.* **9**, 2170–2180 (2000).
15. Blaszczyk, J., Li, Y., Yan, H. & Ji, X. Crystal structure of unligated guanylate kinase from yeast reveals GMP-induced conformational changes. *J. Mol. Biol.* **307**, 247–257 (2001).
16. Sheng, M. & Pak, D. T. Ligand-gated ion channel interactions with cytoskeletal and signaling proteins. *Annu. Rev. Physiol.* **62**, 755–778 (2000).

17. Beguin, P. *et al.* Regulation of Ca^{2+} channel expression at the cell surface by the small G-protein kir/Gem. *Nature* **411**, 701–706 (2001).
18. Finlin, B. S., Crump, S. M., Satin, J. & Andres, D. A. Regulation of voltage-gated calcium channel activity by the Rem and Rad GTPases. *Proc. Natl Acad. Sci. USA* **100**, 14469–14474 (2003).
19. McGee, A. W. *et al.* Structure of the SH3-guanylate kinase module from PSD-95 suggests a mechanism for regulated assembly of MAGUK scaffolding proteins. *Mol. Cell* **8**, 1291–1301 (2001).
20. Garcia, E. P. *et al.* SAP90 binds and clusters kainate receptors causing incomplete desensitization. *Neuron* **21**, 727–739 (1998).
21. Maximov, A., Sudhof, T. C. & Bezprozvanny, I. Association of neuronal calcium channels with modular adaptor proteins. *J. Biol. Chem.* **274**, 24453–24456 (1999).
22. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
23. Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**, 849–861 (1999).
24. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
25. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
26. Jogl, G., Tao, X., Xu, Y. & Tong, L. COMO: a program for combined molecular replacement. *Acta Crystallogr. D* **57**, 1127–1134 (2001).
27. Wu, L., Bauer, C. S., Zhen, X. G., Xie, C. & Yang, J. Dual regulation of voltage-gated calcium channels by $\text{PtdIns}(4,5)\text{P}_2$. *Nature* **419**, 947–952 (2002).
28. Carson, M. Ribbon models of macromolecules. *J. Mol. Graph.* **5**, 103–106 (1987).
29. Nicholls, A., Sharp, K. & Honig, B. H. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* **11**, 281–296 (1991).
30. Kraulis, P. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank Y. Mori for Cav2.1 complementary DNA; E. Perez-Reyes for all β -subunit cDNAs; T. Tanabe for $\alpha 2\delta$ cDNA; and the staff at X4A of NSLS, Brookhaven National Laboratory, for synchrotron support. This work was supported by grants to L.T. and J.Y. from the National Institutes of Health. J.Y. is a recipient of the McKnight Scholar Award, the Scholar Research Programme of the EJB Foundation and the Established Investigator Award of the American Heart Association.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.Y. (jy160@columbia.edu). Coordinates have been deposited in the Protein Data Bank under accession codes 1VYU, 1VYT and 1VYV.

corrigendum

Photonic structures in biology

Pete Vukusic & J. Roy Sambles

Nature **424**, 852–855 (2003).

In this Insight Review Article, a reference to earlier work was inadvertently deleted on proof. The omitted citation is Parker, A. R. *Proc. R. Soc. B* **265**, 967–972 (1998). □

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayouni (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: rasami@naturejp.com

naturejobs

The power of goodbye

Elias Zerhouni, director of the US National Institutes of Health, has come up with a simple way to improve the lot of postdocs. It doesn't involve bureaucracy or legislation or extra pay — just more communication and a refusal by postdocs to accept bad situations. "The best thing is if the brightest mentees walk away from places that don't have good mentoring," Zerhouni told *Naturejobs* (for a complete transcript, see the *Naturejobs* graduate channel, www.nature.com/naturejobs/channels/graduate).

Of course, Zerhouni isn't suggesting that postdocs quit their labs en masse in fits of pique. Rather, he says that before postdocs accept their fellowships, they should sit down with their prospective principal investigator and negotiate their mutual expectations. They can then draw up a sort of scientific "prenuptial agreement", Zerhouni says.

The understanding they reach should include agreement on how much career guidance the mentor will provide, as well as the amount of training that they can expect in such areas as presentation and grant-writing skills. Zerhouni adds that every postdoc will want something different from their mentor. Some prefer a hands-on approach with frequent reporting, whereas others may want a laissez-faire approach.

Zerhouni has a point. Postdocs have power in numbers. Their fellowships mark the last time in their career at which the supply-and-demand formula works in their favour. There are plenty of positions available, especially for postgrads who already have a good track record. Because postdocs need good publications and recommendations from a principal investigator to take the next step, it's essential that they find someone who will help them. And as Zerhouni says, if the principal investigator isn't willing to help, they should find another who will.

Paul Smaglik
Naturejobs editor



Contents

CAREERS AND RECRUITMENT

Molecular diagnostics set to expand p682

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Testing new ground

Diagnosis has traditionally been overshadowed by new drug discoveries. But the expansion of molecular diagnostics is creating new careers for researchers in many fields. Hannah Hoag reports.

Every six months, the software programmers at the molecular-diagnostics business unit of Bayer Diagnostics go back to the drawing board. The programmers responsible for modifying the treatment algorithms for the company's HIV-1-resistance software must update their algorithms regularly because the virus mutates — and acquires drug resistance — so quickly.

The programmers are just one of the groups of professionals Bayer employs to keep its tests up to date. HIV clinicians and researchers regularly meet to share scientific data, dissect high-profile papers and define the impact of new mutations on the virus's resistance to current therapies. At the end of their meetings, the programmers receive a new set of rules, which they translate into new algorithms, transforming the molecular analysis of the virus and its treatment resistance into clinical information that is useful to doctors caring for patients. By the end of the journey, the molecular data have come full circle, back into the hands of the doctors on the frontline of treatment.

Like genomics and functional genomics, molecular diagnostics relies on a network of scientists to make sense of this data-rich field. In addition to programmers, companies well-positioned to make an impact in the molecular-diagnostics market are employing molecular biologists, chemists, biostatisticians and even engineers to bring a test from the research and development (R&D) phase and release it into this fast-growing market.

DETECTION AND PREDICTION

In 1985, the US Food and Drug Administration (FDA) approved a nucleic acid-based diagnostic test for legionnaire's disease from San Diego-based Gene-Probe. Most of the tests that followed in the 1980s and

early 1990s targeted infectious diseases, and tests for HIV, hepatitis C and human papilloma virus (HPV) remain in strong demand today. Lately, molecular tests that diagnose and predict cancers have become more profitable targets. Cardiovascular disease has become another popular target as rates of heart disease increase.

The HPV test marked a shift in the definition of diagnostics, because it not only identified the infection, but predicted a woman's risk of developing cervical cancer. Now the field of molecular diagnostic has developed to include prognostics and biomarkers. The shift from simple detection to prediction has broadened companies' needs. Companies designing prognostic tests that use DNA microarrays may have a greater need for engineers, whereas those that sift through massive amounts of data linking genotype to treatment outcome require biostatisticians and researchers with clinical backgrounds.

GROWING FAST

Historically, the diagnostics industry has been overshadowed by the excitement surrounding new drugs. But times are changing, says Joffre Baker, chief scientific officer at Genomic Health in Redwood City, California. "There's an enormous nascent industry here," he says. "Genomics, sequencing and technologies allowing functional genomics have changed everything."

The US market for molecular diagnostics generated revenues of \$1.3 billion in 2000 and was predicted to generate about \$4.2 billion by 2007 (see *Nature Biotechnol.* **20**, 6–7; 2002). And this means employment. Roche Diagnostics, based in Basel, Switzerland, expects to recruit 250–300 graduates over the next year. Roche is not alone in diagnostic growth. In 1998, Bayer acquired the molecular division of Chiron diagnostics and began its foray into molecular diagnostics. Since then, Bayer's molecular diagnostic tests for HIV and hepatitis C have secured a major share of the market.

Peter Knueppel, head of molecular diagnostics at Bayer Diagnostics in Tarrytown, New York, says that although Bayer's molecular-diagnostics business is not currently looking urgently for staff, the company always has jobs for strong candidates. Molecular biologists are in highest demand, but chemists and software engineers are an important part of the mix. Apart from education, he says, "you need to have a high degree of flexibility and be willing to move quickly — it's a fast-moving industry".

Currently, a shift in disease targets is causing much of the movement within the industry. With plenty of molecular tests available for infectious diseases, companies are moving on to oncology and

Strong alliances

Partnerships are rife in the molecular-diagnostics market. "These are very collaborative days," says Janet Warrington of Affymetrix, which has been actively developing partnerships with Roche Diagnostics to produce a DNA microarray that identifies single-nucleotide polymorphisms in P450 that influence the metabolism of certain psychiatric drugs, and with bioMerieux for HIV genotyping and microbial contamination.

"One company can't do it all," says Peter Knueppel, head of molecular diagnostics at Bayer Diagnostics, who points out that Bayer has partnerships with Genaissance, Innogenetics and GE Healthcare. The strategic alliance between Celera Diagnostics in Alameda, California, and Abbott Laboratories in Abbott Park, Illinois, could cut into some of Roche's market share. Celera's genetic knowledge and Abbott's history of product development will help them produce and market molecular tests for viral and genetic diseases. **H.H.**



AFFYMETRIX

Affymetrix is collaborating with Roche to develop its microarrays for diagnostic detection of single-nucleotide polymorphisms.

cardiovascular-disease diagnostics. In 2001, Abbott Laboratories of Abbott Park, Illinois, acquired the diagnostic company Vysis and began to expand into the oncology arena. In 2002, Abbott began a partnership with Celera Diagnostics of Alameda, California, which is likely to help it secure a larger slice of the molecular-diagnostics business (see 'Strong alliances', box).

TALENT SEARCH

"We're looking for PhDs with industry experience in molecular biology, chemistry and engineering," says Ed Michael, president of Abbott's molecular-diagnostics business unit. Michael says that the molecular-diagnostics division has 40 positions open now, ranging from research to marketing. The company's clinical-research group needs statisticians for number crunching and analysis, and graduates with industry experience could find themselves walking into a senior position at one of Abbott's two campuses near Chicago.

Researchers at Genomic Health are betting their future on molecular tests for cancer. The company, which was spun off from the genomic-information business Incyte, launched its first commercial diagnostic test in February. The 21-gene test predicts the risk of recurrence of certain types of breast cancer. "The test helps identify patients who have a low risk of recurrence and keep them from being exposed to unnecessary chemotherapy," says Baker.

In addition to diagnostics and prognostics, molecular tests are being used to discover patients' responses to treatment. Roche is among the leading companies moving beyond the one-size-fits-all approach to medicine. Last June, in a partnership with Affymetrix of Santa Clara, California, the company launched a DNA chip-based assay test that will identify polymorphisms in two of the drug-metabolizing genes, *CYP2D6* and *CYP2C19*. The test was designed to help doctors choose the best medicine and dosage for their patients, to improve treatment and avoid

some of the 2 million adverse drug reactions that occur in the United States every year and the 100,000 deaths they cause.

The value of genomics may be realized much faster in diagnostics than in therapeutics, but it is still not an easy feat. "Some of the publications coming out of laboratories are nice, but the clinical relevance is sometimes questionable," says Janet Warrington, vice-president of clinical and applied genomics R&D at Affymetrix. "They may be discovering unique biology or pathway information, but it is not an easy step to go from there to a new test." Cost is also an issue. Molecular-diagnostics tests are expensive and in the United States many are not covered by healthcare insurance.

But in the long term, says Baker, the predictive nature of these tests could have a huge economic impact, as they could lead to earlier, less expensive treatments.

As the field grows, companies are employing recent graduates in positions from research to manufacturing to marketing. Industrial or clinical experience is usually essential, but collaboration with other sciences is a priority. "The ideal job candidates are people who have a lifelong interest in this area and who have the ability to pick up information that is outside your primary discipline," says Warrington. Biochemistry, genetics and chemistry graduates predominate in her research group, she says. "They must have some facility using current analytical packages and be able to communicate."

Although molecular biologists predominate in the industry, engineers are equally important, says Schewe. "People don't readily imagine us as an organization recruiting engineers, but it's logical. The instrumentation can be very challenging." Roche's global organization also requires people with worldly skills — including language proficiency and cultural sensitivity. "It's not the scientific expertise alone any more that makes things work. Success or failure depends on how people work together," says Schewe. ■

Hannah Hoag is a science writer in Montreal, Canada.